Branching Process Models of T Cell Selection

Simon Peter Moon

Thesis submitted for the degree of
Doctor of Philosophy

Centre for Mathematics and Physics in the Life Sciences
and Experimental Biology (CoMPLEX)

University College London

July 10, 2006

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Declaration

I, Simon Peter Moon, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

I cannot begin to describe the degree to which I am indebted to my supervisors Andrew George and Jaroslav Stark. I am eternally grateful for their support, tolerance, wisdom and guidance over the last three years. Their patience over this time has been nothing short of saintly. I also wish to thank Robin Callard who also played a significant part in proceedings.
Further gratitude must go to Cliburn Chan whose understanding, help and support I could not have done without. Thanks also to Andy Yates who provided some stimulating and useful information. In addition, I would also like to thank Mary Ritter for her useful input.
Thanks to all at CoMPLEX for their support and help. In particular I would like to thank Rachel and Hugh for their good humour and assistance in administrative matters. I started at CoMPLEX with a remarkable group of people: Dave Dale, Christian Bottomley, Dan Brewer, Olivier Cinquin and Chris Mullaley. I thank them for their comradeship. Particular mention goes to Dan Brewer whose friendship, assistance and ability to organize a night out has been invaluable.
Thanks also to the members of Andrew George's lab at Hammersmith for their comradeship. In particular I would like to thank them for patiently listening to my attempts to explain my work and allowing me to sit in on their lab meetings.
Finally I gratefully acknowledge the MRC for providing the funding required for me to carry out this research.
This thesis is dedicated to my wife Diana whose emotional support has been nothing short of phenomenal. Special mention has to go to my son Sam and my daughter Saskia for making me happy. They are a gift.

Abstract

Cell division history can be examined with the fluorescent dye: 5- (and 6-) carboxyflourescein diacetate succinimidyl ester (CFSE). Modelling the data produced by flow cytometric analysis of CFSE has recently been an active area of research and promises to give new insight into the behaviour of dividing populations of cells. Initially, this thesis describes how CFSE data can be modelled with discrete time multi-type branching processes. This approach has not previously been used where CFSE data is concerned, although branching process models of cell division have an established history of producing simple tractable models that yield biologically relevant results. In particular, these models can be adapted to staged behaviour. We therefore use a multi-type model in an attempt to isolate the phenotypic stage at which negative selection occurs in the thymus. In doing so we re-examine published experimental data that analyses the fate of positively selected double positive (DP CD69$^+$) thymocytes during and after their transition to single positive (SP) stage. We analyse the data with respect to two alternative hypotheses: 1. Death occurs at the DP CD69$^+$ stage and not at the SP stage and 2. Death occurs at the SP stage and not at the DP CD69$^+$ stage and it occurs concurrently with division. We conclude that the second model fits the data better than the first. Motivated to avoid the discrete time assumption that division behaviour is synchronous, the thesis shows how a continuous time branching process model of CFSE can be obtained. The results of a subsequent re-analysis of the published data conflict with our discrete time modelling. Upon further investigation, we conclude that the continuous time model is a poorer model of the data. Finally, the effect of negative selection in combination with division on the thymocyte repertoire is modelled with a discrete time branching process. The results of our analysis suggest that there may be an advantage to division and selection being a combined process.

Contents

List of Figures

List of Tables

List of Abbreviations

APC Antigen presenting cell
CD Cluster of differentiation
CD69 Early activation marker
cdf Cumulative distribution function
CFSE 5- (and 6-) carboxyflourescein diacetate succinimidyl ester
CI Confidence interval
DC Dendritic cell
DP Double positive - refers to thymocytes expressing both CD4 and CD8
DP CD69$^+$ Double positive cell that expresses CD69
FTOC Fetal thymic organ culture
MHC Major histocompatibility complex
MLE Maximum likelihood estimate
pdf Probability density function
pgf Probability generating function
RHS Right hand side
RTOC Re-aggregate thymic organ culture
SP Single positive - refers to a cell expressing only CD4 or CD8
SP CD4$^+$ CD8$^-$ Single positive cell that expresses only CD4
SP CD4$^-$ CD8$^+$ Single positive cell that expresses only CD8
TCR T cell receptor
UGM Unconstrained general model

Chapter 1

Introduction

1.1 Overview

The balance between cell death and division is an important theme in modern biology. The development and maintenance of multi-cellular organisms depends upon the interplay of these opposing processes. For example, during its development the human hand starts life as a fingerless paddle. The digits only emerge later through the controlled and regulated use of cell death (Baehrecke, 2002). Recently our understanding of such processes has advanced with the discovery that the majority of cells destined to die undergo a highly orchestrated sequence of events. This programmed cell death, or apoptosis, involves membrane blebbing, shrinkage, protein fragmentation, chromatin condensation, DNA degradation and engulfment by surrounding cells (Lawen, 2003). Evidence of the importance of balancing death and division is seen when the regulation of either or both mechanisms fails. Pathologies associated with cell overproduction often result, giving rise to developmental defects or cancer.

As with the human hand, the developmental strategy of selecting cells from an excess is echoed in the nervous and immune systems. In the former system, large numbers of cells are destroyed because they fail to make synaptic contacts. In the latter, over production and deletion occurs during the development of T cells in the thymus and this is where our interest lies.

Mature T cells circulate the body and are able to recognize and respond to foreign invaders through their receptors (TCR) (see Chapter 2 for more detailed biological background). These receptors recognize antigens, small peptides derived from pathogens, that are bound to Major Hisocompatibility Complex molecules (MHC) presented on the surface of antigen presenting cells (APC). There is a large diversity of pathogens and consequently a large variety of antigens that are presented. The

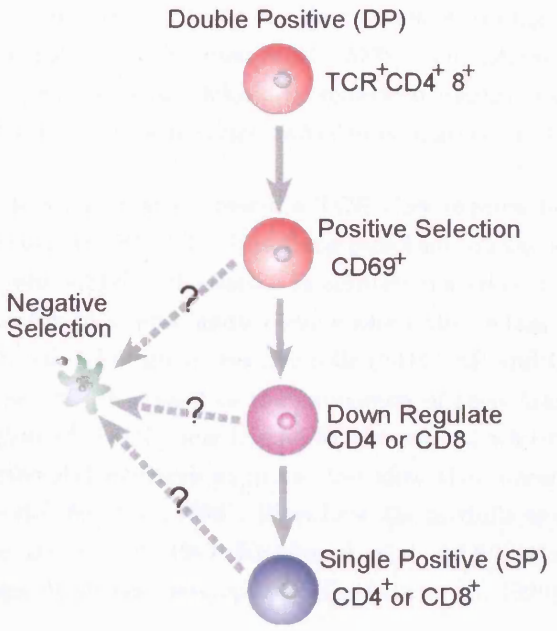

Figure 1.1: Double positive cells express CD4 and CD8 (DP cells). Positive selection is indicated by the expression of CD69. Following positive selection a DP cell down regulates either CD4 or CD8. This results in cells becoming either CD4 or CD8 single positive (SP) cells. The timing of negative selection with respect to the expression of these markers has been a matter of debate.

immune system responds to the challenge of recognizing antigen diversity by creating a repertoire of T cells through variation in their TCRs. This repertoire undergoes a process of education in the thymus.

This education is required because, in addition to presenting foreign antigen, MHC present antigen derived from the self. During their development thymocytes are exposed to these self antigens. The first step of the education process tests whether the repertoire of thymocytes can recognize MHC/self peptides through their TCR. Some fail to recognize anything at all. These useless thymocytes undergo apoptosis. The remaining useful cells are said to have undergone "positive selection". However, this is not the end of the process since some of these positively selected cells recognize self antigen with such strong affinity that they are capable of self-reacting. This possible source of auto-immunity is avoided by a further "negative selection" of these auto-reactive cells. Ultimately, the combined processes of positive and negative selection result in the "central selection" of a repertoire of T cells that upon leaving the thymus are useful and safe (Sebzda et al., 1999).

Attempts to observe and understand the mechanisms that result in the process of central selection

have resulted in some debate. In particular, the developmental timing and location of selection has been controversial (Palmer, 2003; Hogquist et al., 2005). Thymocytes go through various stages during their maturation and these are defined by molecular markers expressed on their surface. In this thesis, we concentrate on stages in which two of these markers ie. CD4 and CD8 are expressed.

At the time when thymocytes first express the TCR they express both CD4 and CD8 and are referred to as double positive (DP) cells. It is these cells that are the subject of positive selection. Upon being positively selected DP cells express an additional marker: CD69. In addition, positively selected DP cells transition to a more mature stage where they either express either CD4 or CD8 alone. This yields CD4 and CD8 single positive cells ($CD4^+$ SP and $CD8^+$ SP) (figure 1.1). The thymic location of these cells is related to the expression of their markers. Double positive cells reside in an outer region of the thymus known as the cortex, whilst SP cells occupy the inner medulla. Some experimental evidence supports the view that negative selection occurs in the cortex at the DP stage (Laufer et al., 1996). Elsewhere, the medulla and its SP occupants are given precedence (Kishimoto and Sprent, 1997; Kishimoto et al., 1998). One group even suggests that selection occurs throughout thymic development (Baldwin et al., 1999).

One aim of this thesis is to attempt, using mathematical modelling, to tackle the problem of identifying the stage at which negative selection occurs. We do so by re-examining data published by Hare et al. (1998). These authors investigated the fate of DP cells that had undergone positive selection as indicated by the expression of CD69 (DP $CD69^+$). They did this without the use of mathematical modelling or any particular interest in cell death. Their method involved the use of 5- (and 6-) carboxyflourescein diacetate succinimidyl ester (CFSE); a dye that enables the experimentalist to track cells through up to 10 divisions (Lyons, 1999, 2000). Their results show that after transition to the SP stage, thymocytes divide. Indeed, in common with Hare et al. (1998), a number of authors have observed SP cell division (Ernst et al., 1995; Penit and Vasseur, 1997; Le Campion et al., 2000, 2002). The presence of division at the SP stage therefore means that we also consider an additional hypothesis: at the SP stage selection and division occur simultaneously.

Given the nature of the CFSE data published by Hare et al. (1998), we use branching process models to analyse the behaviour of their cultured thymocytes (see Chapter 2 for general background on branching processes). Branching process models have a history of being used to model cell behaviour (Jagers, 1975). In addition, they are capable of adaptation to staged behaviour patterns. We therefore create a general discrete time model of the DP to SP transition process and constrain this to test two hypotheses:

1. **Death occurs at the DP $CD69^+$stage and not at the SP stage.**
2. **Death occurs at the SP stage and not at the DP $CD69^+$ stage. If death occurs at this stage it occurs concurrently with division.**

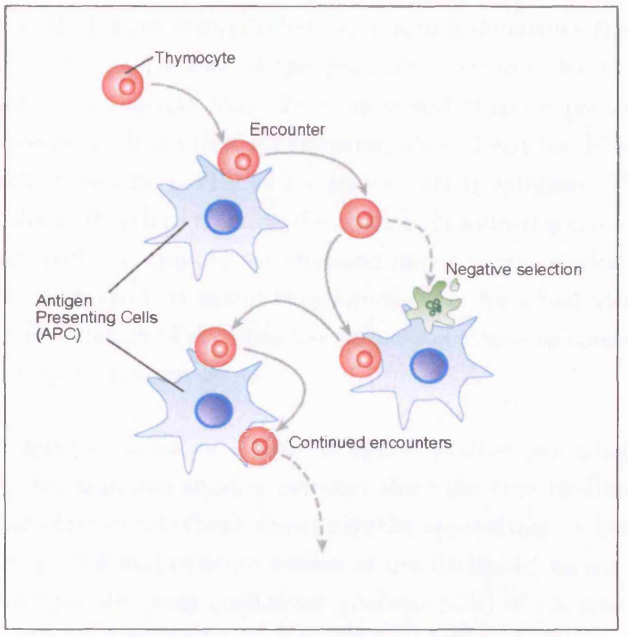

Figure 1.2: Thymocytes undergo multiple encounters with antigen presenting cells (APC). If during one of these encounters the thymocyte recognizes its cognate ligand it is negatively selected and undergoes apoptosis. Otherwise, it continues to sample until it is ready to leave the thymus.

In terms of modelling negative selection our work is unique in two respects. Firstly, it differs from previous mathematical modelling of negative selection in that this has not generally been concerned with the timing and location debate. For example, Mehr et al. (1995) produce a staged model, but make assumptions about the stage at which selection takes place. Elsewhere, selection is typically modelled as being independent of the stage of development (Borghans et al., 1999). Usually, selection is viewed as a sampling process involving a sequence of encounters with cells in the thymic environment (figure 1.2). If during one of these encounters a cell meets its cognate ligand it dies. Such a process is inherently stochastic (Nossal, 1994). Therefore, models have tended to concentrate on either quantifying the extent of negative selection or how such a process can be made safe (Van Den Berg et al., 2001).

Such models generally help support the view that negative selection as described above can work (but see Muller and Bonhoeffer (2003)). However, results take little account of the observation that some auto-immune cells escape negative selection (Bouneaud et al., 2000). Moreover, outside of the thymus there are several mechanisms that deal with auto-reactive cells. Indeed, the induction of this "peripheral tolerance" is an active area of investigation (Lechler et al., 2001). Furthermore, there may be a requirement for some regulatory T cells (Treg) to be self-reactive (Kronenberg and Rudensky, 2005).

Secondly, our inclusion of division during selection is also a departure from previous theoretical work in this area. This may be because of the generally accepted idea that the aim of negative selection is to eliminate auto-reactive cells. Division would create a problem here because when a T cell carrying a specific TCR divides, its offspring also inherit its TCR. This means that all the progeny of a T cell inherit its ability to recognize certain antigens. This clonal propagation of antigen specificity would therefore increase the chances of auto-reactive cells surviving negative selection; although a parent cell may die its offspring may survive or vice versa. Intuitively, the combination of division and selection would therefore seem to be a bad idea. However, in view of the less than complete elimination of auto-reactive cells we continue to consider the possibility that division and selection may be concurrent.

Our models are fitted using a maximum likelihood approximation (see Chapter 2 for the basics of maximum likelihood). An approximation is required since the true likelihood is computationally intractable. A significant part of this thesis shows how the approximation can be applied to test our hypotheses. In particular, the approximate nature of the likelihood means that it is not possible to use standard methods for obtaining confidence intervals (CIs) of parameter estimates. Nor is it possible to follow the standard method for conducting a likelihood ratio test when comparing the fits of our models. Instead, we use Monte-Carlo methods to obtain distributions of estimates and approximate likelihood ratios from which we obtain confidence limits and a test of comparative fit respectively. The results of our analysis suggest that death in these cultures appears to occur at the SP stage. This also suggests therefore that death and division are occurring simultaneously.

Discrete time branching process models assume that cells behave in a synchronous fashion. Motivated by a desire to avoid this assumption, we produce a continuous time branching process model. We also use this to test our hypotheses on the data published by Hare et al. (1998). The result of this analysis suggests that death occurs at the DP stage. This result contradicts the results obtained through the discrete model. We therefore investigate why this is so. The results of our investigation suggest that, in comparison to the discrete model, the continuous time model is a poorer model of the data.

Our results show that division and selection may take place simultaneously, we therefore examine what effect this might have on the quality of negative selection. We do this by using a discrete time branching process model applied to a population of thymocytes. The result can be interpreted in two ways and this is dependent on whether the immune system works on the number or frequency of T cells. Regardless of interpretation the results suggest that there may be an advantage to the combining of division and selection.

1.2 Chapter Outlines

Chapter 2. Biological and Mathematical Background: This chapter contains a basic "text book" introduction to the immunology required to understand this thesis. Some background information on CFSE is also present. In addition, the basic mathematical principles of discrete time simple and multitype branching processes are explained. We also include an introduction to maximum likelihood estimation with reference to the binomial and multinomial distributions. In addition we give an explanation of the standard method of obtaining confidence intervals for maximum likelihood estimates.

Chapter 3. Modelling CFSE Data: the discrete case: Here we demonstrate how to derive a multitype branching process model of CFSE data. We also derive its maximum likelihood function, or estimator, and see that, for large numbers of cells, this function is computationally intractable. Subsequently, we show how the likelihood can be approximated with the use of the multinomial distribution. We also show that when applied to the multinomial approximation the standard method for obtaining 95% CIs for our estimates gives incorrect results. We therefore provide an alternative Monte-Carlo method for obtaining CIs.

Chapter 4. An Application of the Discrete Time Model This chapter contains the formulation of our general discrete branching process model based on the experimental method of Hare et al. (1998). We explain how we constrain the general model to produce 2 models that enable us to test hypotheses 1 and 2 above. We also describe how we use Monte-Carlo methods to produce a likelihood ratio significance test. Our results suggest that, in accordance with hypothesis 2, death occurs at the SP stage. We also test the robustness of our results to sample size. The significance of our results with respect to negative selection are discussed. We conclude that negative selection can occur concurrently with division.

Chapter 5. The Continuous Time Model: We derive a continuous time branching process model based on the assumption of exponentially distributed lifetimes. We use similar methods to those used for the discrete model to analyse the data published by Hare et al. (1998). We find that under continuous time regime our results suggest that, in accordance with hypothesis 1, death occurs at the DP stage. This result is in contradiction to the results of the discrete time modelling. We investigate why this is so. We conclude that our discrete time modelling provides a superior model of the data.

Chapter 6. The Effect of Division and Selection: Here we investigate the effect of division on selection. We also model this using a discrete time branching process. We show how this is applied to a population of thymocytes. In particular we examine how the results can be interpreted depending on whether the immune system operates on the number or frequency of thymocytes. We

conclude that regardless of interpretation there may be an advantage to division during selection.

Chapter 7. Conclusions: We undertake a review of our conclusions from previous chapters. The options for future work are also discussed.

Chapter 2

Biological and Mathematical Background

2.1 Pre-requisite knowledge

An attempt has been made to minimize the amount of pre-requisite knowledge required to understand this thesis. However, it is impossible not to expect some level of prior education. From the biological perspective, some understanding of basic text book cell biology and terminology would be useful. Mathematically, an attempt has also been made to make the thesis accessible to the non-mathematician. Indeed, the work in this thesis requires little more than a basic understanding of probability and matrix algebra. Some understanding of the concept of maximum likelihood would be useful, although a brief explanation is included. In some cases, clarity of explanation has required the loss of some technical rigor. In particular, the explanation of recognition in the immune system is deliberately directed to the non-expert and may to the expert seem superficial.

2.2 Introduction to adaptive immunity, the T cell and the thymus

For the most part the principal reference for this section is Janeway et al. (2001).

Vertebrates continuously come into contact with many types of pathogens such as bacteria and viruses. As a matter of self preservation, these must be identified and subsequently dealt with. The solution that has evolved to answer this challenge is the immune system. This is a highly complex system involving the interaction of many different types of cell. However, we can broadly view the

operation of the immune system as a process of recognition followed by appropriate response. By appropriate we mean that the response is suited to the magnitude and nature of the attack. The recognition step is therefore key to the process because it enables the system to "understand" the infection. Indeed, infections are dynamic and through recognition the system is able to keep track and counter their progression.

Recognition, however, is not a straightforward task since the system must be able to identify non-self from self. Broadly speaking the immune system as a whole has adopted two differing strategies for solving self/non-self recognition. This has led immunologists to view the system as consisting of two separate arms: innate and adaptive[1]. The innate immune system acts as a first line of defense and it is generally characterised by the rapidity and inflammatory nature of its response. In this arm there are various strategies for recognition of non-self. For example, cellular receptors known as pattern-recognition molecules are able to recognize the orientation and spacing of sugar residues which reside on the coatings of certain microbes.

There are a wide variety of pathogens and the innate system does not recognize them all. A further mechanism that has therefore evolved as a second line of defence and this is the adaptive arm of the immune system. Here, recognition is the result of two interacting components: antingen presentation and recognition. An antigen in this case is a short length of peptide (an epitope) and this is sometimes derived from an evolutionarily conserved region of pathogen protein (Hughes and Hughes, 1995). The processing of antigens and their presentation depends upon whether the invader is an intra or extra-cellular pathogen.

In the extra-cellular case the antigen presenting cell (APC) repeatedly samples the external environment through the process of endocytosis. This is an envelopment of a portion of external medium by the cell and results in an internalized body called an endosome. This organelle creates an acidic environment in which pathogens or pathogenic particles are broken down to produce small peptides. Endosomes then fuse with vesicles that contain the molecules whose purpose it is to present extra-cellular antigen at the cell surface: Class II Major Histocompatibility Molecules (MHC Class II). These molecules are membrane bound and consist of two polymorphic polypeptide chains: an α and a β chain (figure 2.1). The conformation of the chains produces a peptide binding groove that both chains contribute to. This groove can hold peptides between 12 and 20 amino acid residues long. After the fusion of the vesicles, the antigenic peptides bind to the binding grooves in the MHC. Once formed the class II MHC/peptide complex is transported to and displayed on the APC's surface.

In the intracellular case there is an added complication. Viruses use the host DNA replication mechanisms to facilitate their reproduction. The process of presentation involves the cleavage of

[1]This simplification is somewhat artificial since there is considerable interaction between the two systems. Indeed, some types of cell play a significant role in both systems.

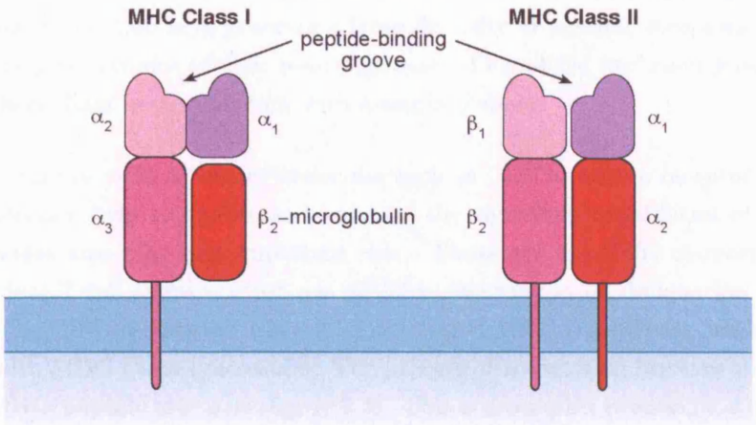

Figure 2.1: MHC class I and class II are membrane bound molecules. MHC class I consists of a polymorphic α chain in conjunction with β_2-microglobulin. The peptide binding groove is formed by the α chain alone. MHC class II molecules consist of polymorphic α and β chains and both these chains contribute to the peptide binding groove. Note that both molecules contain domains named α_1 and α_2 and these are not identical.

protein products by a body called the proteasome. The resulting cleaved peptide fragments are transported into the endoplasmic reticulum where they bind to MHC class I molecules. Unlike MHC class II molecules which are expressed mainly on macrophages etc. MHC class I molecules are expressed on all nucleated cells. This is simply because of the potential for all nucleated cells to be virally infected. MHC class I molecules also differ from class II molecules in that they consist of a heavy polymorphic α chain in combination with the invariant light chain: β_2-microglobulin (figure 2.1). In this case, the peptide binding groove is supplied by the α chain alone and this usually can take antigenic peptides of 9 amino-acid residues in length. Following its formation, the class I MHC/peptide complex is also transported in vesicles to the cell surface.

The cells responsible for recognizing MHC presented antigen and responding to it are T and B cells. This thesis is solely involved with T cell recognition and will therefore ignore B cells. The molecule that enables T cells to recognize the MHC/peptide complex is the T cell receptor (TCR). The majority of mature T cells express a TCR that consists of an α and a β chain, although some express δ and γ chains instead. In this thesis we shall concentrate on describing the processes connected the production of $\alpha\beta$ TCR since the production of $\delta\gamma$ TCR is similar. The α and β chains have variable and constant regions and it is the variable regions which are responsible for antigen recognition. The genes for these peptide chains are found as multiple gene segments in unlinked groups. The α chain variable region is coded for by V (variable) and J (joining) segments: V_α and J_α. There are between 70-80 V_α and 61 J_α segements. To create the variable α chain a V_α and a J_α are randomly recombined. The full α chain is formed when the product of this recombination is conjoined with the constant α chain domain. The TCR β chain is similarly formed through a

process of random recombination, with the addition of a third D (diversity) segement. The effect of this somatic recombination is to generate a large diversity of possible receptors. The expression of the constant region excludes further rearrangement. This allelic exclusion means that the cell expresses a single α chain in combination with a single β chain.

TCR molecules combine with accessory molecules such as CD3 to form a receptor complex (figure 2.2). These molecules help to enable and enhance the signalling capabilities of the TCR. Two additional molecules also play and important role. These are the TCR co-receptors: CD4 and CD8. An individual T cell expresses just one of these, resulting in its designation as either a CD4 or CD8 T cell. The CD4 co-receptor is involved in Class II MHC recognition, whilst CD8 operates in conjunction with MHC Class I molecules. The process of recognition involves the docking of the TCR with the MHC/peptide molecule (figure 2.3). This is a complex process, and investigating the its details is an active area of research (Garcia and Adams, 2005). It is believed that if a TCR has sufficient affinity for an MHC/peptide combination, a signalling process is triggered which causes the T cell to be activated.

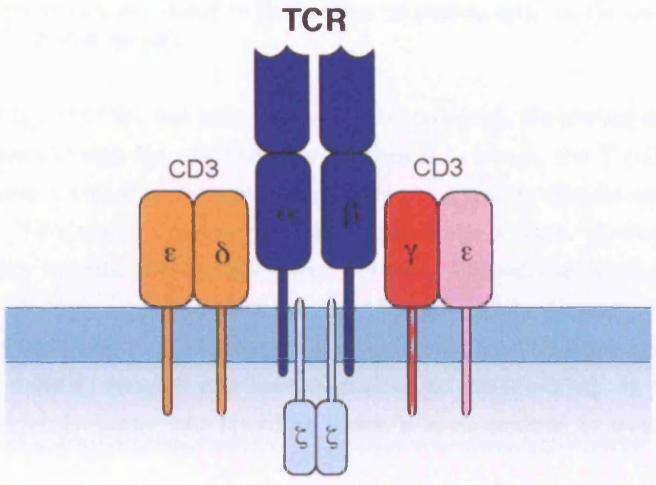

Figure 2.2: The TCR Complex. The TCR consists of an α and β chain and is accompanied by accessory molecules such as CD3. These additional molecules facilitate the TCR's signalling capability.

Both CD4 and CD8 T cells activate and respond to stimulation in a similar fashion by undergoing a clonal expansion. However, their responses differ depending upon co-receptor expression. CD4 cells release cytokines that enhance the activity of other immune cells. On the other hand, CD8 T cells, whose job it is to hunt out virally infected cells, respond by directly killing the cell bearing the viral antigen. Thus CD4 and CD8 T cells are known as helper and cytotoxic T cells respectively.

In summary, we can view the adaptive immune response as a search for a key to a lock. There

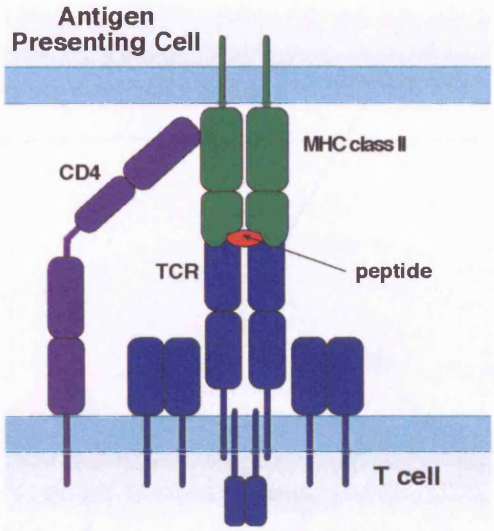

Figure 2.3: The TCR complex recognizes and responds to class II MHC/peptide complex with the aid of CD4. This co-receptor binds to a site distal to the antigen presenting site. In the case of MHC class I the co-receptor involved is CD8 (not shown).

are a large number of keys (TCR) but only one lock (the antigen). Following infection, the system searches through its keys to find the one that fits[2]. When it is found, the T cell bearing this key is activated and undergoes a clonal expansion, increasing the number of cells available to recognize and fight the disease. This whole process runs on a time scale of days. However once the correct key is found it is highly specific for the presented antigen. Indeed two distinguishing features of the adaptive arm are its slow response time and its high specificity for antigens. A third feature of the adaptive arm is its memory. Following an immune response the population of the expanded clone is reduced but remains in higher numbers than prior to the infection. In this way if the same antigen is encountered in the future the speed at which it is recognized is greatly increased.

The methods of antigen processing and presentation give rise to a problem. In regard to Class II presentation the extra-cellular milieu contains host derived peptides and cell fragments. Also, where Class I is concerned the proteasome cannot discriminate between host and pathogenic proteins. This means that inevitably there is presentation of self peptides. Indeed a large proportion of peptides presented on MHC are self derived. Through its TCR, a T cell is only capable of identifying and responding to its cognate antigen and in doing so it also does not differentiate between host and pathogen. This means that the system has to find some way of avoiding T cells responding to self presentation. The solution here is that the generated repertoire of T cells is "educated" by a process of selection and this occurs in the thymus.

[2]The notion that a TCR can only recognize one antigen has been challenged eg. Mason (1998).

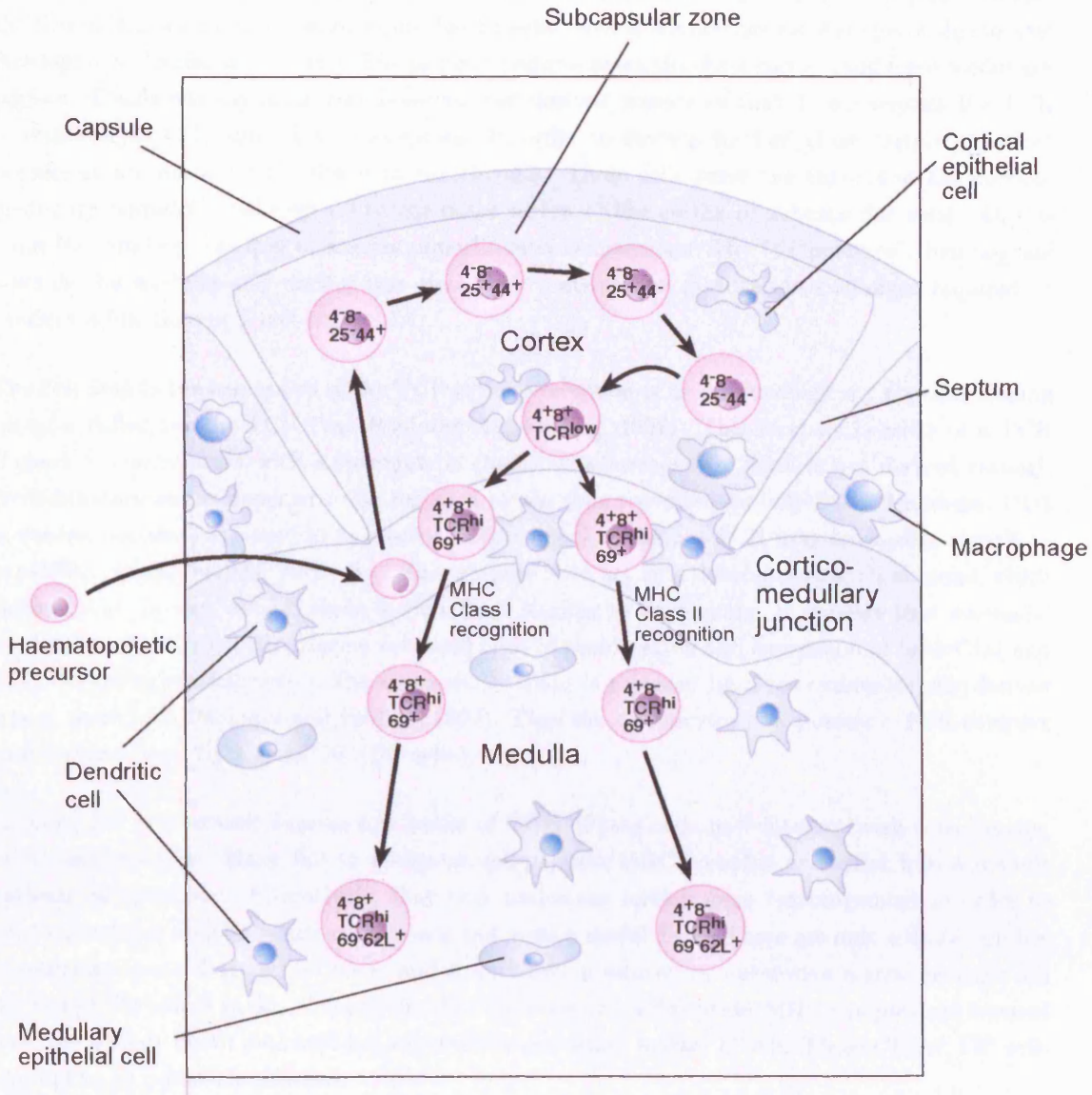

Figure 2.4: The route taken by a developing thymocyte through the thymus. Blood borne precursors enter the thymus at the cortico-medullary junction and migrate to the subcapsular zone. The thymocytes then journey towards the medulla. During this process they undergo phenotypic changes as identified by the surface expression of CD4, CD8, CD25, CD44, CD69, CD62L and also the TCR. After Anderson and Jenkinson (2001) .

The thymus is the site of T cell development. In mammals, this is a small bilobed organ situated just above the heart. The lobes are subdivided into pseudo-lobules by mesenchymal septae. Overall, the thymus has an internal organization that broadly reflects the maturational stages of thymocytic development (Laufer et al., 1999). The primary features being the outer cortical and inner medullary regions. T cells initially arise from bone marrow derived precursors that do not express the TCR or either of its CD4 and CD8 co-receptors. In order to develop further, these "triple negative" precursors are borne by the blood to the thymus. These cells enter the thymus at the cortico-medullary boundary and migrate to the outer region of the cortex or subcapsular zone. At this point they undergo the first of a series of proliferative expansions. The "thymocytes" then migrate towards the medulla and during this time they undergo the maturational changes required to produce a functioning T cell (figure 2.4).

The first step to the expression of the TCR is the production of an immunologically non-functioning receptor called the pre-TCR (von Boehmer and Fehling, 1997). This receptor consists of a TCR β chain in combination with a surrogate α chain. The surrogate α chain is not derived through recombination and is destined to be replaced by the true recombinatorially derived α chain. CD3 molecules are also expressed at this point suggesting that the pre-TCR may have some signalling capability. It appears the purpose of this receptor is to act as a developmental check point which ensures that the expressed β chain is capable of binding to an α chain. It appears that successful production of the pre-TCR triggers a second bout of proliferation and expression of both CD4 and CD8. When expansion ceases, the surrogate α chain is replaced by a re-combinatorially derived true α chain (von Boehmer and Fehling, 1997). Thus the thymocytes now possess a TCR complex that features both CD4 and CD8 (DP cells).

Initially, DP thymocytes express low levels of TCR. These cells now interact with cells bearing MHC/self peptides. Many fail to recognize self-peptide/MHC complex and enter into a default pathway of apoptosis. Alterntively, they may undertake further gene rearrangement in order to produce another α chain which may or may not yield a useful TCR. There are only a finite number of rearrangements that can be made and if all TCRs produced by successive rearrangements fail the fate of the cell is to die. Those cells that do recognize self-peptide/MHC complex are rescued from the default death pathway and express the activation marker CD69. These $CD69^+$ DP cells are said to be positively selected.

The class of MHC encountered when positively selection occurs has an impact on what happens next to the DP cell. If the TCR was triggered by MHC class II the cell down regulates its CD8 expression and becomes a CD4 single positive (SP) cell. Whilst interaction with MHC class I leads to the down regulation of CD4, resulting in a CD8 single positive cell. The transition to the SP phenotype takes place during migration into the thymic medulla.

In addition to the process of positive selection where thymocytes are selected on the basis of being

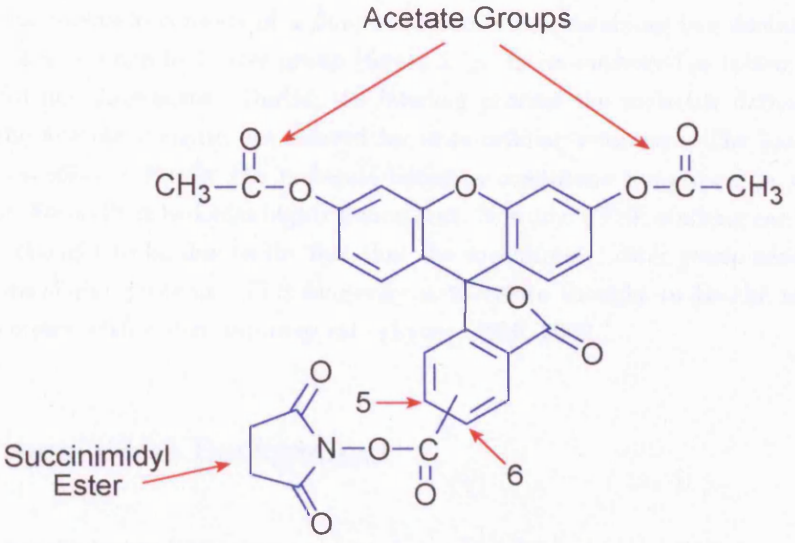

Figure 2.5: The molecular structure of CFSE. The molecule consists of a fluorescein molecule containing two acetate moieties with the addition of a succinimidyl ester group. Here the possible bonding to carbons 5 or 6 is indicated by the bisecting line between them.

useful, another selection process also takes place. This negative selection removes the threat of auto-immunity. It operates on those cells have a strong affinity for self-peptide/MHC complex. Such cells would recognize and respond to the self and are therefore dangerous. The final outcome of both selection processes is the central selection a thymocyte repertoire that is both useful and safe.

2.3 Using 5- (and 6-) carboxyflourescein diacetate succinimidyl ester (CFSE) to track cell division

One approach to observing division in populations of cells is to use CFSE. This dye enables the experimenter to examine the division history of cells. The basic principle of this analysis is, that upon division, the daughters of a cell labeled with CFSE will inherit approximately half the quantity of parentally borne dye each. The loss of fluorescence exhibited by the daughter cells can therefore be directly related to their being the product of a division. Depending on the extent of division up to 10 divisions can be tracked in this way (Lyons, 1999, 2000). Analysis is conducted by flow cytometry and the result is a histogram with distinct peaks each representing populations of cells with identical division histories.

The biochemical properties of CFSE make it particularly useful as a fluorescent marker of division. Stucturally, the molecule consists of a fluorescein molecule containing two acetate moieties with the addition of a succinimidyl ester group (figure 2.5). In its entirety the molecule is membrane permeable and non-fluorescent. During the labeling process the molecule diffuses into the cell, whereupon the acetate moieties are cleaved by intra-cellular esterases. The loss of the acetate groups has two effects. Firstly the molecule becomes membrane impermeable, thus trapping it inside the cell. Secondly it becomes highly fluorescent. Notably, CFSE staining can last for months. This effect is thought to be due to the fact that the succinimidyl ester group binds to free amine groups of intracellular proteins. This longevity is therefore thought to be the result of the dye binding to proteins with a slow turnover rate (Lyons, 1999, 2000).

2.4 Mathematical Background

2.4.1 A simple branching process and probability generating function (pgf)

Let us imagine an individual such an animal or a cell that can reproduce to make other objects like itself. As time passes the individual gives rise to its first generation offspring and these individuals subsequently give rise to further generations. The mathematical tool used to model this type of reproductive phenomena is the branching process. Under specific assumptions, these processes enable us to calculate such things as the probability that, at a some time in the future, the population arising from a single ancestor will consist of a certain number of individuals. Alternatively, using branching processes we can calculate the probability that the entire process will die out. Indeed, branching process models were originally devised in the 19th century to calculate the probability of extinction of aristocratic family names (Harris, 1963; Jagers, 1975).

By way of introduction, here we describe the Galton-Watson process as it applies to cells. For the non-mathematician, this type of branching process is perhaps the easiest to understand. Its primary assumption is that time is divided into discrete generational time steps. Initially, we have a single living cell at generation 0. We imagine that we know the probability that it will either divide (γ), die (α) or remain alive but quiescent (δ), at any given time step (figure 2.6). Since these probabilities are mutually exclusive we have $\gamma + \delta + \alpha = 1$. We also assume that if the cell divides, all its descendants inherit the same probabilities of it undergoing these events. The branching process enables us to calculate the probability that our cell's descendants will number j individuals at generation k defined as

$$P(Z_k = j) = p_{k,j}$$

where j is a non-negative integer. At any given generation k we would therefore have a probability distribution of $p_{k,j}$. For, example given the probabilities γ, δ and α defined above and a starting population of 1 cell ie. $P(Z_0 = 1) = p_{0,1} = 1$ we would have the following distribution at generation 1: $P(Z_1 = 0) = p_{1,0} = \alpha$; $P(Z_1 = 1) = p_{1,1} = \delta$ and $P(Z_1 = 2) = p_{1,2} = \gamma$.

We define the generating function of our probability distribution at generation k as

$$G_k(s) = \sum_{j=0}^{\infty} p_{k,j} s^j \tag{2.1}$$

where the variable s is a dummy variable whose exponent acts as a place marker for the population size (Harris, 1963; Jagers, 1975). To illustrate, we substitute our 3 probabilities into equation 2.1 to obtain at generation 1

$$G_1(s) = \alpha s^0 + \delta s^1 + \gamma s^2 \tag{2.2}$$

It is common practice in the case where $k = 1$ to suppress the suffix on the LHS of equation 2.2 (Jagers, 1975) like so,

$$G(s) = \alpha + \delta s + \gamma s^2 \tag{2.3}$$

and this defines our process generating function. The probability distribution of $p_{k,j}$ at generation 1 is therefore given by the coefficients of the jth powers of s in equation 2.3. If we wish to generate the following generation's probability distribution we now substitute $G(s)$ for s in equation 2.3 (Jagers, 1975). This yields

$$G_2(s) = G(G(s)) = \alpha + \delta(\alpha + \delta s + \gamma s^2) + \gamma(\alpha + \delta s + \gamma s^2)^2$$

and expansion gives,

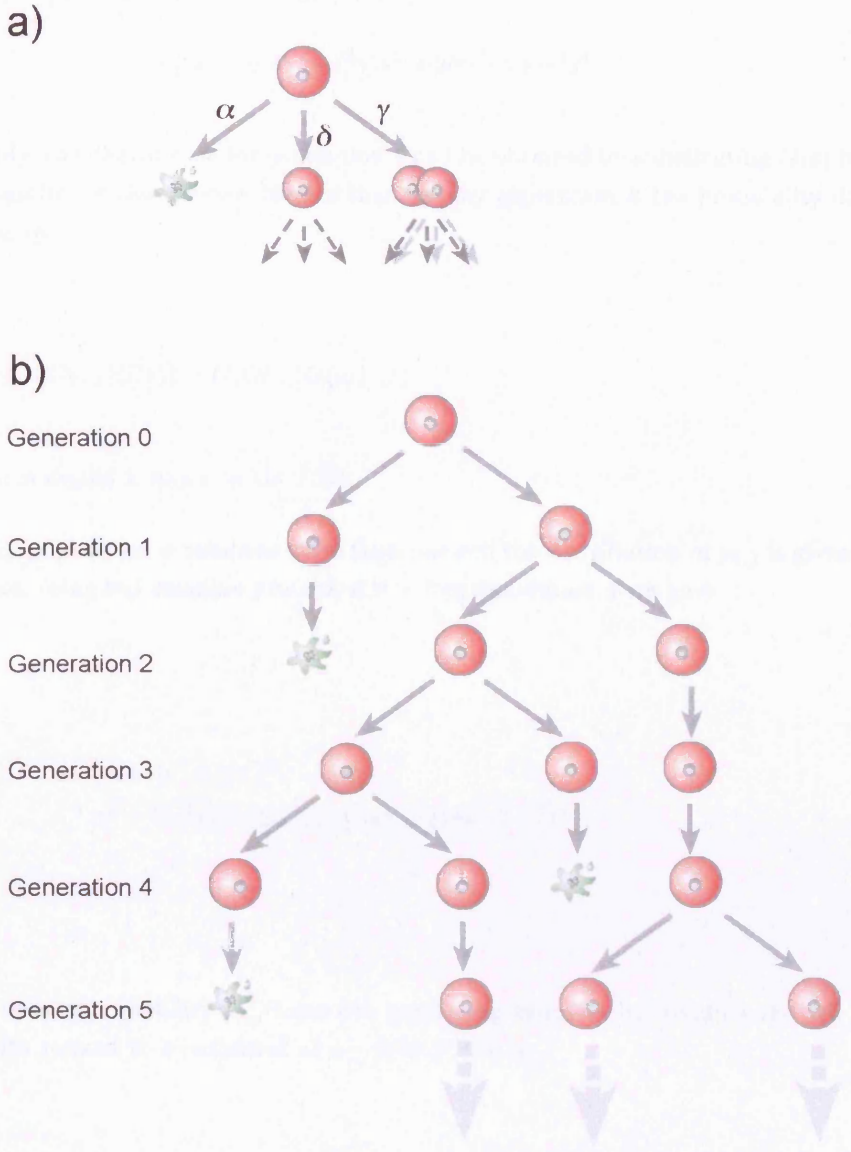

Figure 2.6: The branching process. a) Each cell has 3 distinct fates: division, stasis or death and these depend on the the probabilities γ, δ and $\alpha = 1 - \delta - \gamma$ respectively. b) A typical example of a 3 way branching process including binary division, death and stasis. This process starts with a single cell at generation 0 and by generation 5 has 3 live individuals. After and with thanks to Hardy et al. (2001)

$$G_2(s) = (\alpha + \alpha\delta + \alpha^2\gamma)s^0 + (2\alpha\delta\gamma + \delta^2)s^1$$
$$+(2\alpha\gamma^2 + \delta\gamma + \delta^2\gamma)s^2 + 2\delta\gamma^2 s^3 + \gamma^3 s^4 \tag{2.4}$$

Subsequently, the distribution for generation 3 can be obtained by substituting $G(s)$ for s in $G_2(s)$. Further iteration of this process reveals that for any generation k the probability distribution of $p_{k,j}$ is given by

$$G_k(s) = G_{k-1}(G(s)) = G(G(...(G(s))...))$$

where $G(s)$ is nested k times on the RHS.

If our initial population n numbers more than one cell the distribution of $p_{k,j}$ is given by $[G_k(s)]^n$. For example, using our example process, if $n = 2$ at generation 1 we have

$$\begin{aligned}[G(s)]^2 &= (\alpha + \delta s^1 + \gamma s^2)^2 \\ &= \alpha^2 + 2\alpha\delta s^1 + (2\alpha\gamma + \delta^2)s^2 + 2\delta\gamma s^3 + \gamma^2 s^4\end{aligned}$$

We can recover a probability $p_{k,j}$ from our generating function by dividing the jth derivative of $[G_k(s)]^n$ with respect to s evaluated at $s = 0$ by $j!$ like so

$$p_{k,j} = \frac{1}{j!} \cdot \frac{d^j [G_k(s)]^n}{ds^j}\bigg|_{s=0} \tag{2.5}$$

For example, using our previously defined process with an initial population of $n = 1$, at generation $k = 2$ we recover the probability that there are 3 cells in our process with

$$p_{2,3} = \frac{1}{3!} \cdot \left. \frac{d^3 G_2(s)}{ds^3} \right|_{s=0} = 2\delta\gamma^2$$

Note that from equation 2.5 we obtain the probability of there being zero cells alive at generation k

$$p_{k,0} = [G_k(0)]^n$$

The expectation (the mean number of cells) of a branching process at time step k, denoted by $E[Z_k]$, can also be derived from the generating function (Jagers, 1975). This is obtained from the mean at generation 1 (μ) and this is the first derivative of $G(s)$ (equation 2.3) with respect to s evaluated at $s = 1$ ie.

$$\left. \frac{dG(s)}{ds} \right|_{s=1} = \mu$$

through recursion we find

$$\left. \frac{dG_k(s)}{ds} \right|_{s=1} = \mu^k = E[Z_k]$$

(Harris, 1963; Jagers, 1975). Applied to our simple process this gives

$$\begin{aligned} E[Z_1] &= \delta + 2\gamma \\ E[Z_2] &= (\delta + 2\gamma)^2 \\ E[Z_3] &= (\delta + 2\gamma)^3 \\ &\cdots\cdots \\ E[Z_k] &= (\delta + 2\gamma)^k \end{aligned} \tag{2.6}$$

The variance $V[Z_k]$ of a branching process is obtained from the variance of the process at generation 1 (σ^2)

$$\sigma^2 = \left.\frac{d^2G(s)}{ds^2}\right|_{s=1} + \mu - \mu^2$$

Using recursion the variance at generation k is given by

$$V[Z_k] = \frac{\sigma^2\mu^{k-1}(\mu^k - 1)}{\mu - 1}, \qquad \text{if} \quad \mu \neq 1$$

and

$$V[Z_k] = k\sigma^2, \qquad \text{if} \quad \mu = 1$$

2.4.2 The multitype branching process

We now imagine that we wish to model a population in which there is more than one type of individual. For each parental type i we now define a process probability generating function analogous to equation 2.3 in the previous section

$$G^i(\mathbf{s}) = \sum_{r_1,\ldots,r_j=0}^{\infty} p^i(r_1, \ldots, r_j)s_1^{r_1}, \ldots, s_j^{r_j}$$

where $\mathbf{s}$ is the vector of dummy variables $(s_1, \ldots, s_j)$ that act as place markers for the various types. Whereas, $p^i(r_1, \ldots, r_j)$ is the probability that parent of type i has offspring of types 1to j numbering $r_1, r_2, \ldots, r_j$ (Harris, 1963; Jagers, 1975). For example, in a simple case where there are 2 types ie. $i \in \{1, 2\}$ we have

$$G^i(s_1, s_2) = \sum_{r_1, r_2=0}^{\infty} p^i(r_1, r_2) s_1^{r_1} s_2^{r_2}$$

(see Taneyhill et al. (1999) for a similar example). Returning to the more general case and in a similar manner to the single type branching process at any given time step k we find

$$G_k^i(\mathbf{s}) = G^i[G_{k-1}^1(\mathbf{s}), \ldots, G_{k-1}^j(\mathbf{s})]$$

(Harris, 1963; Jagers, 1975). We can recover the probability that a parent of type i has r_1 offspring of type 1, r_2 offspring of type 2, ...and r_j offspring of type j ie. $p_k^i(r_1, \ldots, r_j)$ from $G_k^i(\mathbf{s})$ by the sequential partial differentiation of $G_k^i(\mathbf{s})$ with respect to s_i. The number of times we differentiate with respect to a given s_i is equal to the value of r_i. The resulting expression is evaluated at all $s_i = 0$ and divided by the product of all $r_i!$ obtaining

$$P_k^i(r_0, r_1, \ldots, r_j) = \frac{1}{\prod_{i=0}^{j} r_i!} \cdot \left. \frac{\partial^{r_0 + r_1 + \cdots + r_j} G_k^i(\mathbf{s})}{\partial s_j^{r_j} \ldots \partial s_1^{r_1} \partial s_0^{r_0}} \right|_{\mathbf{s}=\{0,0,\ldots,0\}} \tag{2.7}$$

Given a vector of initial numbers of individuals at time time step zero $\mathbf{Z}_0$, the expected number of offspring of type j produced by the ith parent at time step k is given by

$$E(\mathbf{Z}_k \mid \mathbf{Z}_0) = \mathbf{Z}_0 \mathbf{M}^k \tag{2.8}$$

where, $\mathbf{M}$ is the matrix of first moments:

$$\mathbf{M} = m_{ij} = E(\mathbf{Z}_1^j \mid \mathbf{Z}_0) = \frac{\partial G^i(1, \ldots, 1)}{\partial s_j} \tag{2.9}$$

(Harris, 1963; Jagers, 1975).

2.4.3 Maximum Likelihood Estimation

The simplest way to explain maximum likelihood estimation is through the example of a coin tossing experiment. We imagine that we will toss a coin n times and that each toss is independent of all others. The probability that in n trials we will produce r heads is given by the binomial distribution:

$$Pr(X = r \mid p) = \binom{n}{r} p^r (1-p)^{n-r} \tag{2.10}$$

where the parameter p is the probability of throwing a head and $1-p$ is the probability of throwing a tail at each toss. The first term on the RHS of equation 2.10 is the binomial coefficient and tells us the number of ways our n trials can be divided into two groups, one containing r heads and the other containing $n - r$ tails. The conditioning on the parameter p on the LHS of equation 2.10 indicates our dependence on the knowledge of the value of p. For example if $p = 0.5$ and we have 10 trials the probability that we will produce 4 heads is

$$Pr(X = r \mid p) = \binom{10}{4} 0.5^4 (1-0.5)^{10-4} \approx 0.205$$

Now we imagine that we have already tossed a coin 10 times and recorded 3 heads and 7 tails. We now wish to calculate an estimate of the value of our parameter p given this data. This is what is called the likelihood and in this case it is given by

$$L(p \mid r) = \binom{10}{3} p^3 (1-p)^{10-3} \tag{2.11}$$

Now in order to find the most likely value of p given the data, or its maximum likelihood estimate (MLE) which we shall denote as $\hat{p}$, we find the maximum of $L(p \mid r)$. Plotting equation 2.11 reveals that $L(p \mid r) = 0.3$ which intuitively makes sense (figure 2.7). An alternative to plotting the likelihood function is to find the maximum through analytical means. This means differentiating $L(p \mid r)$ with respect to p and setting the derivative to equal zero:

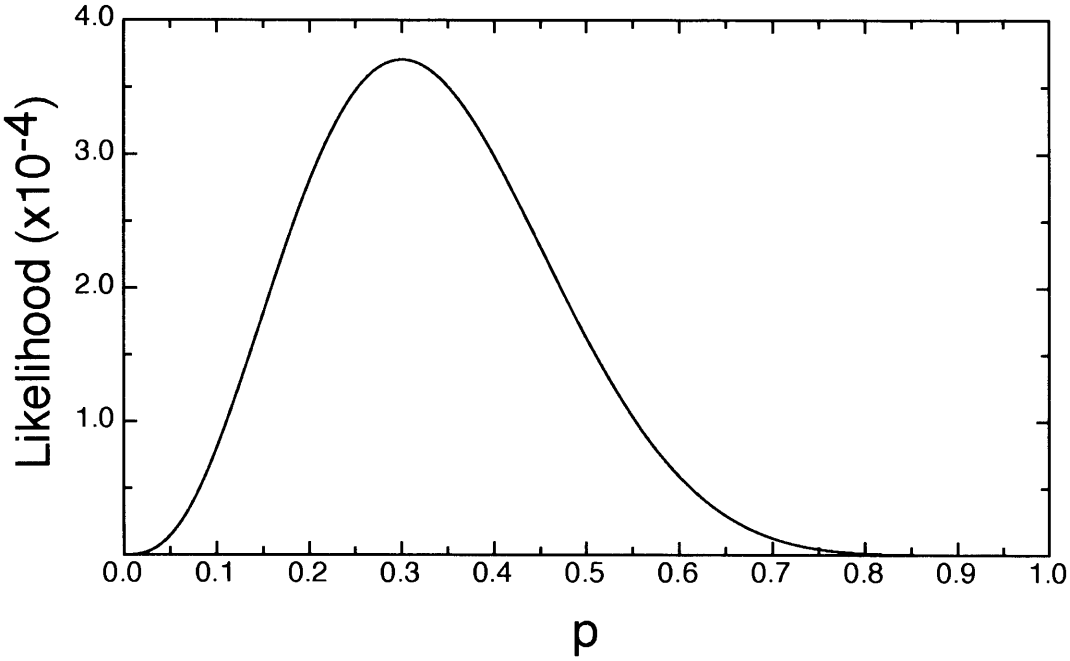

Figure 2.7: The likelihood of a coin tossing experiment that yielded 3 heads and 7 tails. The likelihood reaches a maximum at $p = 0.3$. This is the best estimate of the parameter p = probability of obtaining a head.

$$\frac{dL(p \mid r)}{dp} = \binom{n}{r} \{rp^{r-1}(1-p)^{n-r} - (n-r)p^r(1-p)^{n-r-1}\} = 0 \tag{2.12}$$

substituting $r = 3$ and $n = 10$ and solving for p yields 3 roots: $p = 0$, $p = 1$ and $p = 0.3$. The first 2 of these are minima so our best estimate $\hat{p}$ is 0.3 as above. Strictly speaking we need to check that at this point the second derivative is negative.

In the general case where there is more than 1 parameter we denote the set of parameters as $\boldsymbol{\theta}$ and our likelihood as $L(\boldsymbol{\theta} \mid data)$. Analytically we find the maximum for all parameters θ_i by obtaining and solving the set of simultaneous equations

$$\frac{L(\theta_i \mid data)}{d\theta_i} = 0 \tag{2.13}$$

In many cases the likelihood function is replaced by the computationally more amenable log-

likelihood function $\ell(\boldsymbol{\theta} \mid data)$ since $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta} \mid data)$ also maximizes $\ell(\boldsymbol{\theta} \mid data)$. In cases where solving equation 2.13 is intractable it is necessary to resort to numerical methods for finding maxima such as the Nelder-Mead simplex algorithm (Press et al., 2002).

2.4.4 Maximum Likelihood and the multinomial distribution

We have seen how, given 2 mutually exclusive outcomes, the binomial distribution gives us the probability of there being r successes in n trials. If we wish to model the scenario where there are more than 2 mutually exclusive outcomes to choose from, we use the multinomial distribution. Given $k+1$ possible outcomes this is

$$Pr(X_1 = r_1; X_2 = r_2; \ldots; X_k = r_k \mid p_i, p_2, \ldots, p_k) = \frac{n!}{r_1! r_2! \ldots r_{k+1}!} p_1^{r_1} p_2^{r_2} \ldots p_{k+1}^{r_{k+1}} \tag{2.14}$$

where r_i indicates the number of times outcome i has occurred in all n trials. The probability of outcome i occurring at each trial is given by p_i. Note that since all outcomes are mutually exclusive $r_{k+1} = n - r_1 - r_2 - \ldots - r_k$. By the same token the sum of all p_i is unity and therefore $p_{k+1} = 1 - p_1 - p_2 - \ldots - p_k$. So in order to calculate a multinomial probability we only require k parameters. The first term on the RHS side is the multinomial coefficient and tells us the number of ways that we can obtain r_1 of outcome 1 and r_2 of outcome 2 etc. in n trials.

As an example, let us imagine that we have 3 possible outcomes at each trial ie. $i \in \{1, 2, 3\}$ with probabilities $p_1 = .3, p_2 = .5$ and $p_3 = 1 - p_1 - p_2 = 1 - .3 - .5 = .2$. Given this information, we wish to calculate the probability that in $n = 10$ trials outcomes 1 and 2 occur 2 and 4 times respectively. Remember the number of times outcome $k+1 = 3$ occurs is given by $r_3 = n - r_1 - r_2 = 10 - 2 - 4 = 4$. Substitution into equation 2.14 yields

$$Pr(X_1 = 2; X_2 = 4 \mid .3, .5) = \frac{10!}{2!4!4!} .3^2 .5^4 .2^4 = .02835$$

If we are given data that tells us how many of each outcome occurred during n trials we can obtain the likelihood $L(\boldsymbol{\theta} \mid data)$ of the parameters $\boldsymbol{\theta} = \{p_1, p_2, \ldots, p_k\}$. The likelihood does not depend upon the multinomial coefficient and therefore this can be dropped from our calculations. Thus the multinomial likelihood is given by

$$L(\boldsymbol{\theta} \mid X_1 = r_1; X_2 = r_2; \ldots; X_k = r_k) = \prod_{i=1}^{k+1} p_i^{r_i} \tag{2.15}$$

taking logs we obtain the log-likelihood

$$\ell(\boldsymbol{\theta} \mid data) = \sum_{i=1}^{k+1} r_i \log[p_i] \tag{2.16}$$

As discussed above solving the system of simultaneous equations 2.13 obtains the MLEs of our parameters.

2.4.5 Confidence limits for MLEs

If we were to undertake multiple repeats of an experiment we would not expect all data sets obtained to be identical. In the example of our 10 tosses of coin, this would be true even if we use the same coin throughout. If we took the data produced by such multiple experiments and found the maximum value of the likelihood function for each data set we would obtain distributions of both likelihoods and MLEs. The best estimate of a particular parameter would be described by its distribution mean and 95% confidence interval (CI). The standard method for finding the confidence interval for MLEs, without repeating the experiment multiple times, is to approximate the likelihood distribution with a gaussian distribution. This is achieved by taking the Taylor expansion of the likelihood function $\ell(\boldsymbol{\theta} \mid data)$ about its maximum. In the one parameter case, we denote the MLE of our parameter p as $\hat{p}$ and also $\ell = \ell(p) = \ell(p \mid data)$ and expansion yields

$$\ell = \ell(\hat{p}) + \frac{1}{2} \left. \frac{d^2\ell}{dp^2} \right|_{\hat{p}} (p - \hat{p})^2 + \ldots \tag{2.17}$$

Since $\hat{p}$ is found at

$$\left. \frac{d\ell}{dp} \right|_{\hat{p}} = 0$$

the first order term of the expansion is zero. The first term in the Taylor series 2.17 is a constant and does not provide any information about the shape of ℓ. Ignoring greater than second order terms, the quadratic term therefore dominates in determining the variance of ℓ. The exponential form of series 2.17 therefore yields

$$\ell \propto A \exp\left(\frac{1}{2}\frac{d^2\ell}{dp^2}\bigg|_{\hat{p}} (p-\hat{p})^2\right) \tag{2.18}$$

where A is a constant. Equation 2.18 has a form similar to the normal or gaussian distribution:

$$Pr(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.19}$$

The standard deviation $sd = \sigma$ of ℓ is subsequently given by the square root of the negative inverse of the second derivative:

$$\sigma = \left(-\frac{d^2\ell}{dp^2}\bigg|_{\hat{p}}\right)^{-1/2}$$

In the multivariate case with k parameters we denote the vector of parameters MLEs $\hat{\boldsymbol{\theta}}$. Expansion yields the quadratic approximation to the multivariate normal distribution:

$$\ell \propto \exp\left[\frac{1}{2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^T \mathbf{H}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)\right] \tag{2.20}$$

where $\mathbf{H}$ is the $k \times k$ matrix of second derivatives, or Hessian, with the ijth element $\partial^2\ell/\partial\theta_i\partial\theta_j$. Analogously to the single parameter case, the negative inverse of the Hessian (evaluated at $\hat{\boldsymbol{\theta}}$) is the covariance matrix $\boldsymbol{\sigma}^2$. The square roots of the diagonal ($i = j$) components of the covariance matrix supply the standard deviations of our θ_is. The off-diagonal elements ($i \neq j$) provide information about how parameter θ_i covaries with θ_j. We subsequently will refer to this method of finding CIs as the quadratic approximation.

Chapter 3

Modelling CFSE data: the discrete case

3.1 Introduction

In recent years the modelling of CFSE data has been an active area of research (Bernard et al., 2003; Pilyugin et al., 2003; Leon et al., 2004; Ganusov et al., 2005; Boer and Perelson, 2005; Chan et al., manuscript in preparation). In a purely immunological context modelling has successfully been applied to the behaviour of cells outside the thymus (Gett and Hodgkin, 2000; Hodgkin, 2005). These latter authors are of particular interest because their method requires the re-scaling, or normalization, of the data. Their method has the effect of transforming the data from cell numbers to quantities of CFSE dye (also see Pilyugin et al. (2003)). Here we take a similar approach by re-scaling the data and modelling the division distribution of CFSE dye.

First we explain how to use a multitype branching process to model CFSE cell count data. We see that for large cell numbers this leads to a computationally intractable likelihood. We subsequently show how we can use the multinomial distribution to construct an approximate likelihood function. Based on the re-scaling of the expectations of cell numbers, predicted by the multitype branching process, this function allows us to obtain parameter estimates from dye distributions derived from re-scaled cell count data.

We test the multinomial approximation using a Monte-Carlo method. Our results show that, given data generated with known parameter values, the multinomial approximation produces estimates that are close to these parameters (the estimator is approximately unbiased). We also show that the standard method of quadratic approximation for obtaining CIs does not work for the multinomial

approximation. The correct 95% CIs must therefore be produced by our Monte-Carlo method. We mention two alternative methods for obtaining the likelihood: quasi-likelihood and normal approximation. We argue that the multinomial approximation has an advantage over these methods in its speed and flexibility.

3.2 The multitype branching process and division history

The multi-type branching process enables us to model data that shows the division history of cells such as that provided by CFSE analysis. Here we will once again use our simple branching process as an example. First we define the type of a cell to be equal to the number of divisions it has undergone. For example a cell that has not divided will be type 0. A cell that has divided once will be type 1. Those that have divided twice will be type 2 and so on. Our probability of division γ is now the probability that parental type i can give birth to 2 offspring of type $i+1$ at each generational time step. The probability of doing nothing δ is the probability that type i will remain type i at each time step. Once again $\alpha = 1 - \delta - \gamma$ is the probability that a cell of type i will die at each time step. From this information, we now define the generating function for a parent of type i

$$G^i(\mathbf{s}) = \alpha + \delta s_i + \gamma s_{i+1}^2 \tag{3.1}$$

If we start with one cell of type 0 its generating function will be:

$$G^0(\mathbf{s}) = \alpha + \delta s_0 + \gamma s_1^2 \tag{3.2}$$

At generation 2 we substitute $G^0(\mathbf{s})$ for s_0 and $G^1(\mathbf{s})$ for s_1 in equation 3.2, yielding

$$G_2^0(\mathbf{s}) = \alpha + \delta(\alpha + \delta s_0 + \gamma s_1^2) + \gamma(\alpha + \delta s_1 + \gamma s_2^2)^2$$

Iteration reveals that in order to obtain $G_k^0(s)$ we undertake the substitution of $G^i(\mathbf{s})$ for each s_i in $G_{k-1}^i(\mathbf{s})$. This therefore models a process where we see the first cells of type k at generation k.

If the number of cells n of type 0 in our initial population is greater than 1 and we wish to find

the probability distribution of types at generation k, we follow the rules applied to the simple one type branching process by raising $G_k^0(\mathbf{s})$ to the nth power. For example if we have 2 cells of type 0 in our initial population then at time step 1 we have

$$\begin{aligned}[G^0(\mathbf{s})]^2 &= (\alpha + \delta s_0 + \gamma s_1)^2 \\ &= \alpha^2 + 2\alpha\delta s_0 + \delta^2 s_0^2 + 2\delta\gamma s_0 s_1^2 + 2\alpha\gamma s_1^2 + \gamma^2 s_1^4\end{aligned}$$

Once again the coefficients of the dummy variables provide their associated probabilities. For example, at generation 1, the probability that there are two cells that have divided once accompanied by 1 cell that has not is given by the coefficient of $s_0 s_1^2$.

Analytically, this can be obtained using an expression similar to equation 2.7. For an initial population of n cells of type 0 and a sequence of the number of cells of each type from 0 to j ie. $(r_0, r_1, \ldots, r_j)$ at time step k we sequentially partially differentiate $[G_k^0(\mathbf{s})]^n$ with respect to s_i. The number of times we differentiate with respect to a given s_i is equal to the value of r_i. The resulting expression is evaluated at all $s_i = 0$ and divided by the product of all $r_i!$ obtaining

$$P_k^0(r_0, r_1, \ldots, r_j) = \frac{1}{\prod_{i=0}^{j} r_i!} \cdot \left. \frac{\partial^{r_0 + r_1 + \cdots + r_j} [G_k^0(\mathbf{s})]^n}{\partial s_j^{r_j} \ldots \partial s_1^{r_1} \partial s_0^{r_0}} \right|_{\mathbf{s}=\{0,0,\ldots,0\}} \tag{3.3}$$

It follows from sections 2.4.3 and 2.4.4 that the likelihood of $\boldsymbol{\theta} = (\delta, \gamma)$ given data $\mathbf{r} = (r_0, r_1, r_2, \ldots)$ ie. $L(\boldsymbol{\theta} \mid \mathbf{r})$ is obtained by maximizing the RHS of equation 3.3[1]. For data consisting of low numbers of cells (see below) maximization of this function is probably achievable. However, for a large initial population n and with large numbers of cells in each division category this expression is computationally intractable. This is because it involves high dimensional nested loops in order to keep track of the high order terms involved. For example, with a simple branching process and a starting population of the order of 10^6 we require at least one loop per generation each of size 10^6. Optimization of the likelihood would involve multiple evaluations of this function and this would only increase the impracticable nature of the problem.

[1]Note that it is not necessary to estimate α since $\alpha = 1 - \delta - \gamma$.

Precisely how many cells in the initial population that can be computationally catered for by equation 3.3 has not been investigated. This is because our aim here is the analysis of data that usually involves numbers of cells of the order of 10^5 to 10^6. However, bearing in mind the comments made above it seems unlikely that current technology could produce a workable maximization program for more than a few hundred cells in the initial population.

3.3 An approximation to the likelihood

CFSE data usually contains large numbers of cells and therefore the intractability of the likelihood means that we must find an alternative approach if we wish to find estimates of the parameters $\boldsymbol{\theta}$. In this section we show that if the number of cells in any given division history category is assumed to be independent of all others, a multinomial likelihood can be used to approximate the true likelihood. The assumption of category independence is not easy to justify since each generation of each process that contributes to the overall distribution of cells is dependent upon its preceding generation. However, as we shall see, in practice the multinomial approximation produces excellent results in terms of providing accurate parameter estimates (see section 3.4.2 and Chapter 4). The reason why the multinomial approximation works remains an unanswered question and is probably the greatest mathematical challenge that can be set in regard to future research in this area.

Our approximation is based on the expected numbers of cells in each division category. It also depends upon switching our perspective from modelling cell numbers to modelling the dye itself. Using the simple branching process as an example, we shall proceed by first explaining each step in deriving the approximate likelihood function.

3.3.1 Step 1: Obtaining the expected numbers of cells in each division category

In section 2.4.2 we saw how given the vector of the initial number of cells of type i ie. $\mathbf{Z}_0$, we obtain the expectation $E(\mathbf{Z}_k \mid \mathbf{Z}_0) = \mathbf{Z}_0\mathbf{M}^k$ of the number of offspring of type j produced by the ith parent at time step k (equations 2.8 and 2.9). Since we are using a simple branching process at time step k we include type $k+1$ since the pgf at time step 1 for types $i = 0, 1, \ldots, k$ from equation 3.1 above is

$$G^i(\mathbf{s}) = \alpha + \delta s_i + \gamma s_{i+1}^2$$

and type $k+1$ has pgf

$$G^{k+1}(\mathbf{s}) = \delta s_{k+1} + (1-\delta)$$

Subsequently $\mathbf{M}$ is a $(k+1)\times(k+1)$ matrix. As an example, in the case where we observe 3 types we have

$$\mathbf{M} = \begin{bmatrix} \delta & 2\gamma & 0 & 0 \\ 0 & \delta & 2\gamma & 0 \\ 0 & 0 & \delta & 2\gamma \\ 0 & 0 & 0 & \delta \end{bmatrix} \tag{3.4}$$

In the simplest case where the process started with 1 cell of type 0 ie. $\mathbf{Z}_0 = \{1,0,0,0\}$ our expectation of the cell numbers in each division category is

$$E(\mathbf{Z}_3|\mathbf{Z}_0) = \mathbf{Z}_0\mathbf{M}^3 = \{\delta^3, 6\delta^2\gamma, 12\gamma^2\delta, 8\gamma^3\} \tag{3.5}$$

Thus the expected number of cells in a given division category $E[r_i]$ is given by the ith element of the RHS of equation 4.5. For example, the expected number of cells that have divided twice is $12\gamma^2\delta$.

3.3.2 Step 2: Modelling the distribution of dye

The relationship between the number of divisions a cell has undergone and the amount of dye it's ancestors contained is straight forward. If a cell in the starting population contains 1 unit of dye then its immediate offspring contain 0.5 units. The second generation contain 0.25 units and so on. Given the expected number of cells in a division category at time step k we can therefore obtain the expected amount of dye $E[d_i]$ it contains from

$$E[d_i] = 2^{-i}E[r_i] \tag{3.6}$$

If we have one cell in the initial population the expected distribution of dye $E[\mathbf{D}_1]$ at generational time steps 1, 2 and 3

$$\begin{aligned} E[\mathbf{D}_1] &= \{\delta, \gamma\} \\ E[\mathbf{D}_2] &= \{\delta^2, \delta\gamma, \gamma^2\} \\ E[\mathbf{D}_3] &= \{\delta^3, 3\delta^2\gamma, 3\gamma^2\delta, \gamma^3\} \end{aligned}$$

by iteration we see that the expected proportion of dye in each division category at generation k is therefore given by the binomial distribution ie.

$$E[d_i] = \binom{k}{i} \gamma^i \delta^{k-i} \tag{3.7}$$

3.3.3 Step 3: The proportion of lost dye

The experimentalist only counts living cells. When a cell dies the amount of dye it contains is lost. The expected proportion of dye seen in equation 3.7 is conditional on the cells being alive and therefore the expected proportion of dye lost $E[\xi_k]$ at generation k will be given by:

$$E[\xi_k] = 1 - \sum_{i=0}^{k} E[d_i] \tag{3.8}$$

3.3.4 Step 4: Transforming the data

We now find the expected amount of dye contained in a division category by normalizing it. Thus, from the number of cells r_i in division category i we obtain the expected amount of dye they contain d_i

$$d_i = 2^{-i} r_i \tag{3.9}$$

Assuming each cell in the initial population n contains one unit of dye, we infer the amount of dye lost through death to be this initial amount minus the sum of all dye found in division categories 0 to k. This yields

$$\xi_k = n - \sum_{i=0}^{k} d_i \tag{3.10}$$

3.3.5 Step 5: The multinomial approximation

In section 2.4.4 we saw how the multinomial distribution models data that contains more than 2 independent categories. If we assume that each division category is independent of all others we can therefore view CFSE data as belonging to a multinomial distribution. This is also based on the assumption that each cell in the initial population n contains one unit of dye. The number of trials here is subsequently given by n. In addition, the probability that a unit of dye will be found in a given category is given by the expectation $E[d_i]$. For the simple branching process this would result in the approximate likelihood function or estimator

$$L(\boldsymbol{\theta} \mid data) = \frac{n!}{d_0! d_1! \ldots d_n! l_k!} E[\xi_k]^{\xi_k} \prod_{i=0}^{k} E[d_i]^{d_i} \tag{3.11}$$

where $\boldsymbol{\theta}$ represents the set of our parameters. As noted in section 2.4.4 we can ignore the multinomial coefficient, since the parameters have no dependency upon it. We therefore derive the computationally more convenient log-likelihood function

$$l(\boldsymbol{\theta} \mid data) = \xi_k \log E[\xi_k] + \sum_{i=0}^{k} d_i \log E[d_i] \tag{3.12}$$

3.4 A test of the multinomial approximation

3.4.1 A Monte-Carlo test

If a maximum likelihood estimator provides an accurate estimate of a parameter it is said to be unbiased (Mood, 1963). Given the approximate nature of our estimator we therefore need to test that the estimates it produces are approximately unbiased. In addition, the fact that our estimator is an approximation means that we must also test whether the CIs produced by the quadratic approximation (section 2.4.5) are acceptable. We therefore test for bias in our estimator and check the accuracy of CIs produced by the quadratic approximation using the following steps:

1. For a given model, we calculate the expected cell numbers at generational time step k.
2. Estimate the parameter values from this data set.
3. Calculate 95% CIs using the quadratic approximation.
4. Use the parameter values and the given model to produce 1000 simulated data sets.
5. Obtain further parameter estimates for each of the simulated data sets. For each parameter we know have distribution of MLEs.
6. The mean of an estimate distribution can be compared to our original parameter value. This gives an indication the bias of our estimate.
7. For a given parameter, rank the simulated estimates and select the 26th and 975th of these as asymmetric 95% CIs.
8. Compare these to the 95% CIs obtained through the quadratic approximation.

3.4.2 Results of the Monte-Carlo test

Following steps 1 to 8 above using arbitrarily chosen parameter values of $k = 5$, $\beta = 0.4$ and $\gamma = 0.25$ our results showed that the multinomial estimator is unbiased (Figure 3.1 and Table 3.1). In figure 3.1 this is indicated by the close correspondence of the input paramater values and the distribution means. In general this result holds for large n (eg. $n \geq 10^5$) (C. Chan personal communication). Further evidence of the unbiased nature of the multinomial estimator is also provided by tables 4.4 to 4.9 in the following chapter.

The true CIs derived through simulation are narrower than those produced by the quadratic approximation. This result may or may not generalize to all combinations of parameter values. In

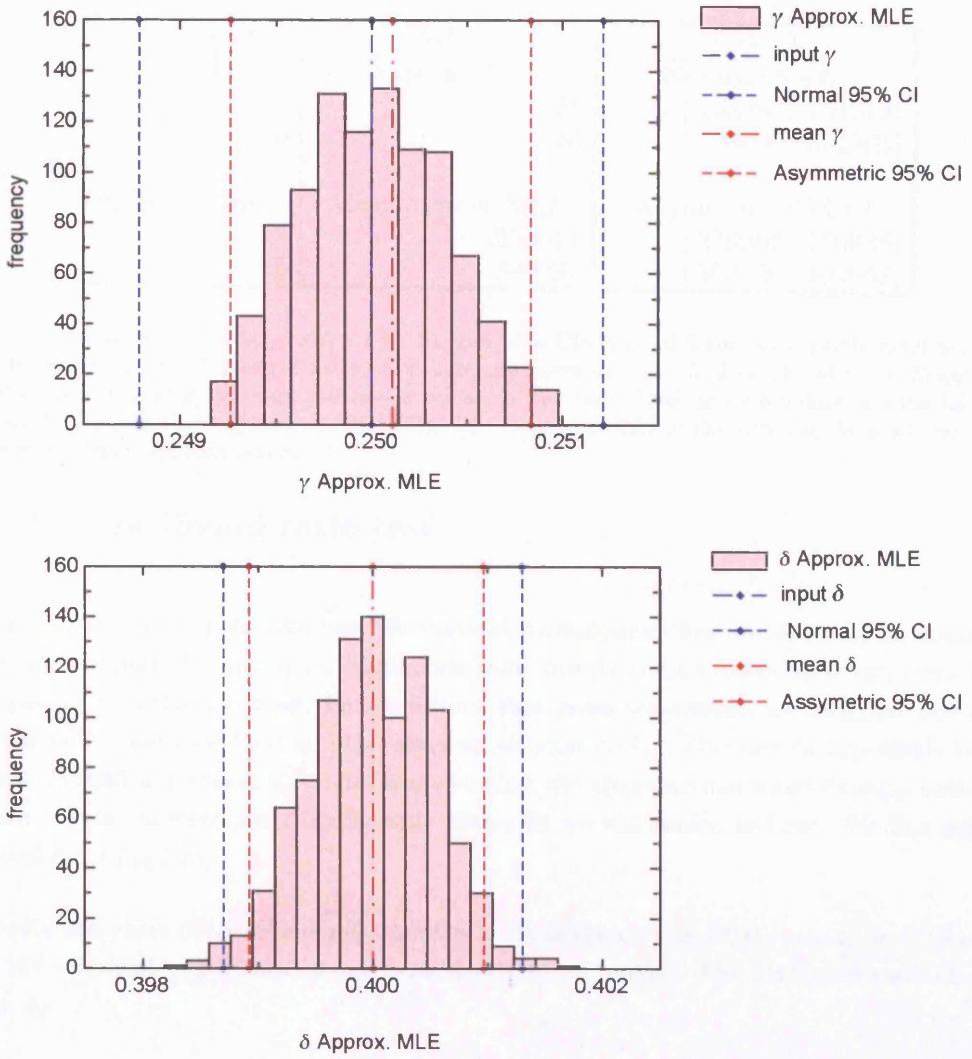

Figure 3.1: The distributions of the approximate MLEs derived from 1000 simulations of a simple discrete branching process with parameters a)$\gamma = 0.25$ and b)$\beta = 0.4$. The number of time steps $k = 5$. The red vertical lines indicate the distribution means and asymmetric 95% CIs. The blue vertical lines indicate the input parameter values and Normal 95% CIs obtained using the quadratic approximation. In the case of β the input parameter value and distribution mean are indistinguishable at this scale.

other regions of parameter space the quadratic approximation may produce CIs that are narrower than or equal to the true CIs. Regardless of this, we argue that this one instance of inaccuracy is enough for us to consider the quadratic approximation as an unreliable method when applied to the multinomial approximation. Subsequently, it was decided that CIs would be produced by Monte-Carlo simulation. This has an added advantage in that it would also produce further evidence that MLEs produced by the multinomial approximation are unbiased.

Parameter	MLE	95% CI
a) Test Data	Approx. MLE	Normal 95% CI
γ	.25	[.248787, .251213]
δ	.40	[.398697, .401303]
b) Simulation	Mean Approx. MLE	Asymmetric 95% CI
γ	.250011	[.249265, .250835]
δ	.399991	[.398920, .400967]

Table 3.1: A comparison of approximate MLEs and 95% CIs derived from: a) a single artificial data set calculated from the expectations of a simple branching process with $k = 5$, $\beta = 0.4$ and $\gamma = 0.25$ and b) 1000 simulated data sets using the same parameter values. In the case of the simulated data sets the table shows the mean MLEs and their asymmetric 95% CIs. The Normal 95%CI of the artificial data set was obtained using the quadratic approximation

3.5 The likelihood ratio test

In isolation, the value of the likelihood function at its maximum has no particular meaning. However, if we compare the maximum likelihoods from two different models their ratio can be used as a basis for hypothesis testing. Let us assume that given two models we designate one of these to be our null hypothesis (H_0) and the other an alternate (H_A). Our aim in hypothesis testing is to reject the null hypothesis if it can be shown that the alternate has a significantly better fit to the data. What we mean by a significantly better fit we will return to later. We first define the likelihood ratio $L_r(data)$.

We denote the value of the likelihood function at its maximum as $L^\star(\boldsymbol{\theta}_0 \mid data)$ or $L^\star(\boldsymbol{\theta}_A \mid data)$ where the subscript of $\boldsymbol{\theta}$ indicates our dependence on the model. The likelihood ratio ($L_r(data)$) is given by

$$L_r(data) = \frac{L^\star(\boldsymbol{\theta}_A \mid data)}{L^\star(\boldsymbol{\theta}_0 \mid data)} \tag{3.13}$$

the log-likelihood difference is often referred to as the log-likelihood ratio (as it will be in this thesis) is subsequently given by

$$\ell_r(data) = \ell^\star(\boldsymbol{\theta}_A \mid data) - \ell^\star(\boldsymbol{\theta}_0 \mid data) \tag{3.14}$$

The likelihood ratio indicates how much better H_A explains or fits the data than H_0. In general, if the log-likelihood $\ell_r(data) > 0$ then H_A is a better explanation of the data than H_0. If $\ell_r(data) = 0$ then both models fit equally well and if $\ell_r(data) < 0$ then H_0 fits the data better than H_A. In the latter two cases we go no further since we are only interested in rejecting H_0 if H_A has a significantly better fit and these results indicate that H_0 is the better fitting model. In the former case we therefore need to define what we mean by a significantly better fit. The likelihood ratio test has the form

$$\text{Reject } H_0 \text{ if } \ell_r(data) \geq \kappa \tag{3.15}$$

where κ is a suitable constant to be decided upon. We therefore reject H_0 if H_A gives us at least κ times as good an explanation of the data. However, there are two types of error that can be made when undertaking significance testing: a Type I error is committed if we reject H_0 when it is true and a Type II error is committed when we fail to reject it if it is false. In terms of the scientific method committing a Type II error is considered the lesser of the two evils. This is because, even if you fail to reject the null hypothesis, it does not mean that you have shown it to be correct. In order to guard against a Type I error we therefore protect H_0 by only rejecting it if the significance level or p value of our test is small. The significance level is defined as the maximum probability at which we are prepared to reject H_0 assuming it is true. Typically, in biology the significance level is set at $p \leq 0.05$.

This all becomes clearer in application. In our particular case, we will be defining two different branching process models: Model 1 and Model 2. These both seek to explain some CFSE data. Maximizing the likelihood on the data for both models provides us with two likelihoods which we compare using equation 3.14. We do this by casting one of our models in the role of the null model and the other as the alternative. Note that the choice of which model will act as a null here is an arbitrary choice between models 1 and 2. This is usually not the case, since a null model typically reflects the possibility that data may have occurred completely by chance. However, given that $\ell_r(data) \neq 0$ one of our models will have a better fit ($\ell_r(data) > 0$) and we therefore wish to test whether this is significant.

As an example we will assume that Model 2 has a better fit to the data than Model 1. Using equation 3.14 we see that there are two possibilities. Firstly, with Model 2 cast in the role of null and Model 1 as the alternate we would find that: $\ell_r(data) < 0$. This result says that Model 2 has a better fit than Model 1 and we would not reject Model 2 if it were our null model. Reversing the roles of our models would mean that $\ell_r(data) > 0$. As we would expect, this would also indicate that Model 2 has a better fit than Model 1. However, because Model 1 is now our null model we

need to know whether Model 2 has a significantly better fit than Model 1 before we can decide whether to reject Model 1 or not.

In this latter case and using equation 3.14 we would have

$$\ell_r(data) = \ell(\boldsymbol{\theta}_2 \mid data) - \ell(\boldsymbol{\theta}_1 \mid data) > 0 \tag{3.16}$$

Where the subscripts now indicate our models 1 and 2. Typically, $2\ell_r(data)$ (known as the Deviance) has a χ^2 distribution. Usually, we can therefore obtain a p value from a χ^2 statistical table. However, in the case of an approximate likelihood this would not be the case as this depends on using the true likelihood. We therefore turn to a Monte-Carlo method of likelihood ratio significance testing. To achieve this we take the following steps:

1. Use the parameter estimates provided by maximizing the likelihood for the null model (in our example this would be Model 1) and generate a number of data sets (usually 1000) through Monte Carlo simulation.
2. Find the maximum likelihoods using both models to produce a pair of likelihoods for each data set.
3. Obtain the likelihood ratio for each data set using its respective pair of likelihoods.
4. The result is a set of likelihood ratios based on the assumption that the null model is correct. To visualize this, we can plot the frequency distribution of these ratios (eg. figure 4.7 in chapter 4).
5. Our original likelihood ratio taken on the experimental data ($\ell_r(data_e)$) can now be compared to our simulated ratio distribution by plotting it (also in figure 4.7). Alternatively, we can obtain CIs for the distribution against which the ratio on the experimental data can be compared.

With respect to this last point, the usual biological significance level of $p \leq 0.05$ we would normally reject the null if $\ell_r(data_e)$ lies outside of the 95% CIs of the ratio distribution. However, given that there is uncertainty about the variance of our estimator we therefore choose only to reject the null (Model 1 in our example) if $p \leq .001$. This means that working with 1000 data sets we would only reject the null model if $\ell_r(data_e)$ lies completely outside of the ratio distribution.

3.6 Alternative Estimators

Two alternative estimators can be obtained using quasi-likelihood and a normal approximation. A normal approximation assumes that the variance of the expected cell numbers has a multivariate normal distribution. The quasi-likelihood is based on the idea that the expectation and its variance are sufficient information from which we can derive an estimator. However, both these methods require the calculation of the process covariance matrix. This is a complicated procedure and not generally a trivial task.

We argue here that the simplicity of the multinomial approximation has distinct advantages over the quasi-likelihood and normal approximations. The chief advantage being that of speed and flexibility. Models can be quickly reformulated using the multinomial approximation whereas the alternatives require complete recalculation of their process covariance matrices.

Chapter 4

An application of the discrete time model

4.1 Introduction

4.1.1 An application of the multinomial approximation

In this chapter we use the multinomial approximation in order to re-examine CFSE data published by Hare et al. (1998). In doing so we enter into a debate relating to the timing and location of negative selection. There are two facets to our investigation. Firstly, we ask whether DP or SP cells are the subject of negative selection. Secondly, since SP cells are known to divide we ask if division and selection could be occurring concurrently. By way of introduction we first examine the literature relating to both selection and cell division during the latter stages of thymocyte development.

4.1.2 Timing and Location of Negative Selection

The timing and location of negative selection has been a matter of much debate (Palmer, 2003). In a recent review Hogquist et al. (2005) state that 32 different transgenic mouse models have been used to investigate negative selection. In experiments, roughly half of these produce data that suggests that negative selection occurs in the cortex (DP selection) and half that it occurs in the medulla (SP selection). Typically, these studies use mice that have thymocytes that are monomorphic for a certain TCR and these are deleted when challenged by MHC bearing the TCRs cognate ligand. An

early example of this approach used a TCR specific for male HY antigen. This antigen is present in male but not in female mice. Subsequently it was demonstrated that, in the male transgenic mouse, exposure to the antigen causes DP thymocytes in the thymic cortex apoptose (Kisielow et al., 1988; von Boehmer, 1990). This result suggested that DP cells are the subject of negative selection. However, the use of transgenic mice may introduce non-physiological artifacts. Often the expression of a TCR clonotype found in a transgenic model is premature or may be expressed at higher levels than we would expect in wild type mice. This may therefore result in the premature deletion of thymocytes (Hogquist et al., 2005; Baldwin et al., 2005).

An example of work supporting the view that the medulla is the site of negative selection is provided by Kishimoto and Sprent (1997). These authors found that semi-mature SP thymocytes, cells that reside in the medulla, are subject to deletion following stimulation using anti-TCR monoclonal antibody (mAb) (Kishimoto and Sprent, 1997). They later support this conclusion using the super-antigen staphyloccal enterotoxin B (SEB) (Kishimoto et al., 1998). In general super antigens, such as SEB, have high affinity for TCR but operate by binding to their V_β domains rather than the peptide binding groove. This means that the deletion generated by super-antigens may differ from self-antigen influenced deletion. Results of this nature and those using mAb therefore suggest that SP cells are capable of being negatively selected rather than giving evidence that they actually are under physiological conditions.

Circumstantial evidence that the medulla is the site of negative selection comes from the fact that this area is enriched with important APCs: dendritic cells (DCs) and medullary epithelial cells (TECs) (Palmer, 2003). However, it has been suggested that expression of certain self-peptides may also be location specific (Hogquist et al., 2005). This would mean that the timing of deletion would be dependent on whether the peptide was expressed in the cortex or medulla. Timing may also be due to the affinity that a TCR has for its cognate ligand (Sant'Angelo and Janeway, 2002). Generally, consensus points to selection occurring in the medulla. This would seem to imply that SP cells are the subject of deletion. However, transition from the DP to the SP phenotype also occurs whilst the cells are migrating through the cortico medullary junction and into the medulla. The exact phenotype of cells subject to negative selection therefore is open to debate.

4.1.3 Division and Selection?

In adult mice low levels of division have been seen at all the phenotypic stages of development that follow expression of the TCR (Penit and Vasseur, 1997). Indeed, positive selection of DP cells is identified by the transient expression of CD69, a marker of activation in the periphery (Lucas et al, 1994). Penit and Vasseur (1997) report that, *in vivo* around 7% of DP cells expressing CD69 incorporate BrdU, an indicator of cell cycling. Lower BrdU incorporation (1- 3% of cells) is

observed in the intermediate phenotypes between DP and mature SP cells. In addition, in mature SP cells division is also observed the data indicating that around 1.7% of $CD4^+CD8^-$ and 5.24% of $CD4^-CD8^+$ cells incorporated BrdU.

It has been observed that peri and neo-natal mice exhibit much higher levels of thymocyte division than adults (Ceredig, 1990; Ernst et al., 1995; Hare et al., 1998). It is thought that this expansion plays an important role in establishing the T cell repertoire (Hare et al., 1998). The observations of Ceredig (1990) also suggested that cycling SP $CD4^+CD8^-$ cells are the subject of negative selection.

In assessing the data presented by many experimentalists in this field, matters are not helped by a lack of uniformity in the delineation of phenotypic groups. Furthermore, the lack of models against which data can be compared has led to a scarcity of statistical analysis in the presentation of data. However, we conclude there is clear evidence that cell division occurs in thymocytes at stages when negative selection is likely to occur and in particular SP cell division is frequently observed. We therefore argue that concurrent selection and division is a possibility that cannot be overlooked.

4.1.4 CFSE analysis in a thymic context

In a thymic context Hare et al. (1998) used CFSE analysis to investigate the fate of DP cells that have undergone positive selection, as indicated by the expression of the activation marker CD69 (DP $CD69^+$). Their basic method was to use re-aggregate thymic organ cultures (RTOCs; reviewed in Hare et al. (1999b)) that were pulsed with CFSE prior to incubation. After an incubation period lasting up to 3 days these cultures were harvested for thymocytes. During this time some of the DP cells had undergone phenotypic transformation into both CD4 and CD8 SP cells. In doing so DP cells were not seen to divide. However, following transition to SP cells division occurred.

Further to the above, the transition to SP cells can occur in the absence of thymic stromal cell support. However, in the absence of stromal cells the subsequent wave of division was not observed. Upon the addition of purified thymic epithelium alone or whole thymic stroma, SP division was observed. Hare et al. (1998) therefore suggest that "thymic epithelium is both necessary and sufficient" to drive the division of SP cells. Moreover, they also produced data which suggests that division may occur in an MHC independent manner. They therefore speculated that the production of cytokines by thymic epithelial cells may be the possible driving component in SP division. In particular epithelial cells express IL-7, a cytokine known to stimulate cell division. Subsequent work by these authors using thymic stroma deficient for MHC (Hare et al., 1999a) and later, thymocytes deficient for the IL7 receptor (IL-7R) (Hare et al., 2000) goes on to investigate this speculation. Here their results suggested that SP division is independent of MHC interaction and mediated by IL-7R.

In their analysis Hare et al. (1998) do not enter into the examination of the timing of cell death in their cultures. This is not an oversight but unnecessary from the point of view of the authors in investigating their specific hypotheses. However, given that we have a devised a method that enables us to model and analyse CFSE data we initially asked whether it would be possible to re-examine the data using a model or models that can determine the timing of cell death in these cultures. Such information may help to shed light on the debate as to exactly when and where negative selection takes place.

Following the approximate maximum likelihood approach, we therefore established a general discrete time branching process model that can be constrained to test two hypotheses:

1. **Death occurs at the DP CD69$^+$ stage and not at the SP stage.**
2. **Death occurs at the SP stage and not at the DP CD69$^+$ stage. If death occurs at this stage it occurs concurrently with division.**

Note that these hypotheses are examined assuming that DP CD69$^+$ cells do not divide. This assumption follows the evidence described in Hare et al. (1998) and elsewhere (Ceredig, 1990; Ernst et al., 1995). Our results suggest that, in these cultures, cells appear not to die at the DP stage. Additionally, since death appears not to occur at the DP stage we suggest that death occurs at the SP stage during a time that these cells are also dividing.

4.2 Mathematical methods

4.2.1 The General Model

The first step in examining the phenotypic timing of cell death is to create a general model that encompasses the possible behaviour of the cells at each of the 2 phenotypic stages. Using discrete time, the cell's behaviour at each time step was therefore assumed to be as follows (figure 4.1)

1. **Assumption 1:** DP CD69$^+$cells have probability β of making the transition to the SP stage. In addition, these cells are able to remain in the DP CD69$^+$stage whilst neither dividing nor dying with probability δ_1. We also assume, in line with the data provided by Hare et al. (1998), that DP CD69$^+$cells do not divide. The probability of death at this stage is given by $\alpha_1 = 1 - \delta_1 - \beta$.

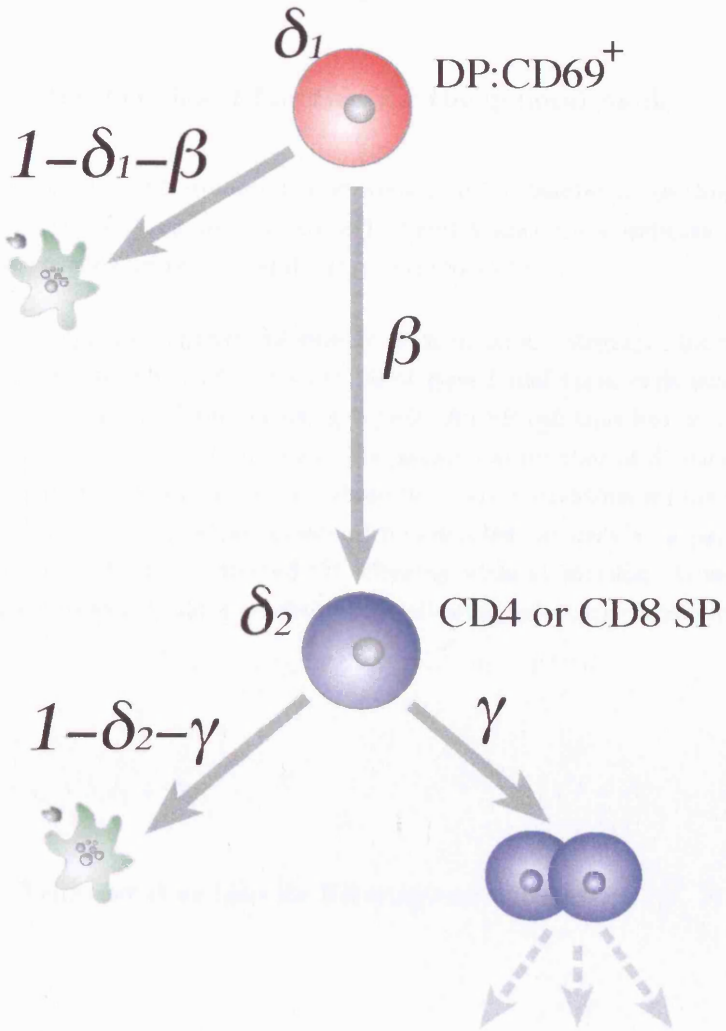

Figure 4.1: The discrete time general model: At each time step, DP CD69$^+$cells can transition to SP cells with probability β or remain as DP CD69$^+$cells without dying with probability δ_1. Their probability of dying is therefore $1-\delta_1-\beta$. SP cells follow a typical 3 way branching process behaviour (see figure 2.6) ie. at each time step, they can divide with probability, do nothing or die with probabilities γ, δ_2 and $1-\delta_2-\gamma$ respectively.

2. **Assumption 2:** Reflecting the irreversible nature of DP to SP transition we modelled SP cells as unable to make a reverse transition to DP cells. However we assume these cells follow a typical 3 way branching process behaviour (figure 2.6) in which cells either divide, die or do nothing with probabilities γ, δ_2 and $\alpha_2 = 1 - \delta_2 - \gamma$ respectively.

4.2.2 Deriving the likelihood function for the general model

Using the assumptions 1 and 2 we now follow steps 1 to 5 (Chapter 3: section 3.3) to derive our likelihood function. We will demonstrate steps 1, 2 and 3 since these indicate the basic difference between the simple branching process and our general model.

First we derive the expected numbers of cells in each division category. For this general model, we define our non-dividing DP CD69$^+$cells to be of type 1 and these cells can transit to not yet divided SP cells which we will define as being type 2. An SP cell that has divided once is defined as type 3, divided twice as type 4 and so on. The maximum number of divisions accounted for in the model is κ. In order to model a process where there are κ divisions we therefore include $\kappa + 2$ types. Here we regard DP CD69$^+$precursors of non-divided SP cells as a parental type, we say that DP CD69$^+$cells produce non-divided SP offspring without division. Given the probabilities outlined in our Assumptions 1 and 2 we obtain the following generating function for our type 1 DP CD69$^+$cells

$$G^1(s_1, s_2) = \alpha_1 + \delta_1 s_1 + \beta s_2 \tag{4.1}$$

for SP cells (type 2 and above) we have the following general equation

$$G^i(s_i, s_{i+1}) = \alpha_2 + \delta_2 s_i + \gamma s_{i+1}^2 \tag{4.2}$$

for $2 \leq i \leq \kappa + 1$. For type $\kappa + 2$ we have

$$G^{\kappa+1}(s_{\kappa+2}) = \delta_2 s_{\kappa+2} + (1 - \delta_2) \tag{4.3}$$

From these we derive the matrix of first moments $\mathbf{M}$. As an example, if we are required to model data that contains cells with up to 2 SP divisions we require 4 types, representing DP cells (type 1) and the SP cells that have divided 0, 1 and 2 times (types 2, 3 and 4 respectively). Given this information we obtain from equation 2.9 and equations 4.1, 4.2 and 4.3 the matrix of first moments

$$\mathbf{M} = \begin{bmatrix} \delta_1 & \beta & 0 & 0 \\ 0 & \delta_2 & 2\gamma & 0 \\ 0 & 0 & \delta_2 & 2\gamma \\ 0 & 0 & 0 & \delta_2 \end{bmatrix} \tag{4.4}$$

Taking the simplest case where $Z_0 = \{1, 0, 0, 0\}$, that is we initially have just 1 DP cell in our population, then the expected numbers of each type at time step 3 are given by

$$E(\mathbf{Z}_3|\mathbf{Z}_0) = \mathbf{Z}_0\mathbf{M}^3 = \{\delta_1^3, \beta(\delta_1^2 + \delta_1\delta_2 + \delta_2^2), 2\beta\gamma(\delta_1 + 2\delta_2), 4\beta\gamma^2\} \tag{4.5}$$

Note that in the case of our general model, we obtain the expectation that there are zero divisions from the sum of the first 2 components of this vector, since type 1 and type 2 represent cells that have not undergone any divisions. Therefore, in order to produce a vector whose ith component is the expectation of the ith division category $E[z_i]$ we take the additional step of multiplying $E(\mathbf{Z}_\kappa|\mathbf{Z}_0)$ by a column vector $\mathbf{A}$ of length $\kappa + 2$ with entries $a_i = 1$ for $i = 1, 2$ and $a_i = 0$ for $2 < i \leq \kappa + 2$.

Thus given the starting populations of each type of cell and the framework of our general model we can derive the expected cell numbers $E[z_i]$ of each division category i at time step k. Working with these expected numbers and we subsequently follow steps 2 to 5 in order to produce our estimator. Using our 2 division example with the maximum number of time steps $k = 3$ we obtain the expected distribution of dye:

$$E[\mathbf{D}_3] = \{\delta_1^3 + \beta(\delta_1^2 + \delta_1\delta_2 + \delta_2^2), \beta\gamma(\delta_1 + 2\delta_2), \beta\gamma^2\} \tag{4.6}$$

and

$$E[l_3] = 1 - (\delta_1^3 + \beta(\delta_1^2 + \delta_1\delta_2 + \delta_2^2) + \beta\gamma(\delta_1 + 2\delta_2) + \beta\gamma^2) \tag{4.7}$$

the expected proportion of dye lost.

4.2.3 The Models

Model 1

We produce a model of hypothesis 1 (hypothesis 1) by constraining our general model's parameters so that

$$\delta_2 + \gamma = 1 \tag{4.8}$$

this means that we create a model where only DP cells are allowed to die (figure 4.2). Since the probability of death at the SP stage is given by $\alpha_2 = 1 - \delta_2 - \gamma$ our constraint means that $\alpha_2 = 0$ ie. SP cells cannot die. Furthermore, this constraint reduces the number of parameters by one as either δ_2 or γ can be expressed in terms of each other. Here we choose to set $\delta_2 = 1 - \gamma$.

Model 2

Based on the notion that there is evidence to support the site of negative selection as the thymic medulla, a site where the vast majority of SP cells reside, we produce a model of hypothesis 2 (model 2) in which we use the constraint

$$\delta_1 + \beta = 1 \tag{4.9}$$

In biological terms this constraint means that DP cells cannot die and can only transit to the SP stage where they are subsequently able to either die, divide or do nothing (figure 4.3). This constraint also reduces the number of parameters in the model since either δ_1 or β can be expressed in terms of each other. Here we set $\delta_1 = 1 - \beta$.

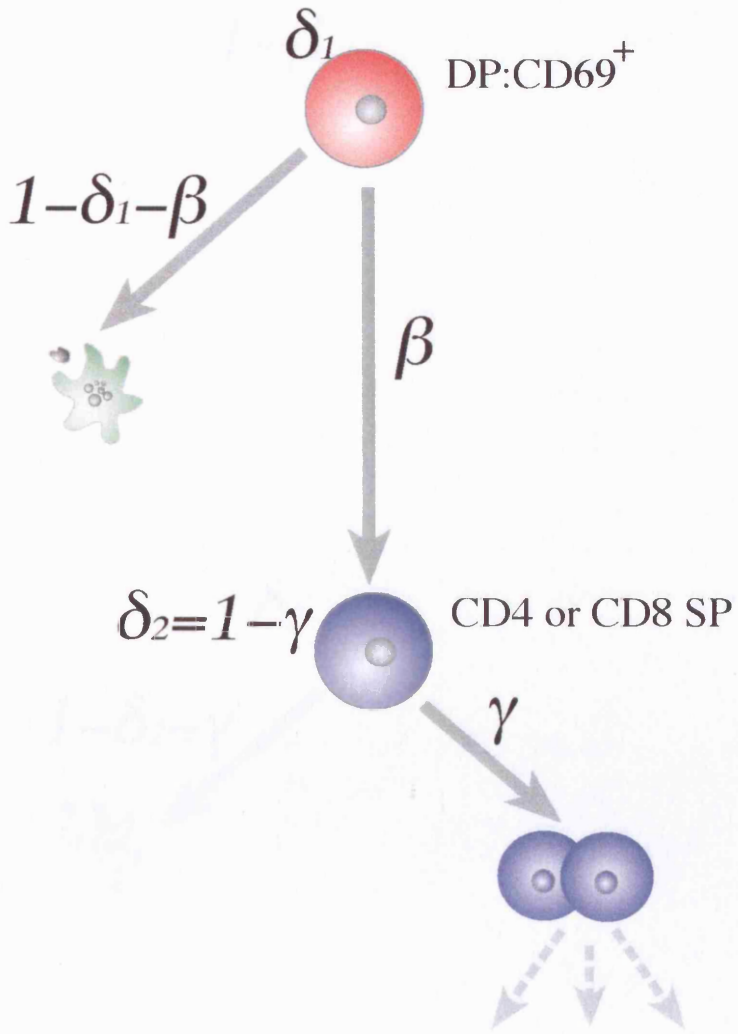

Figure 4.2: Discrete time Model 1: at each time step, DP CD69$^+$ cells can either transit to the SP phenotype, do nothing or die with probabilities β, δ_1 and $1-\delta_1-\beta$ respectively. SP cells can either divide or do nothing with respective probabilities γ and $\delta_2 = 1-\gamma$.

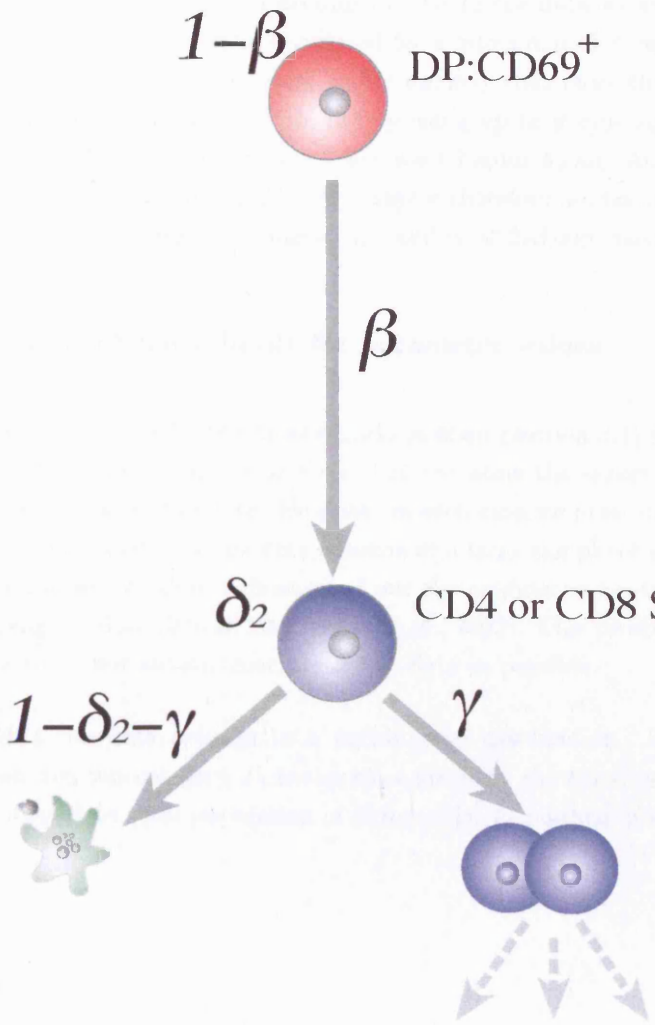

Figure 4.3: Discrete time Model 2: at each time step, DP CD69$^+$ cells can either transit to the SP phenotype or do nothing with probabilities β and $\delta_1 = 1 - \beta$ respectively. SP cells can either divide, do nothing or die with respective probabilities γ, δ_2 and $1 - \delta_2 - \gamma$.

4.2.4 The number of time steps

In modelling the data provided by Hare et al. (1998) we initially set the number of time steps used in our models to be the minimum required to obtain the maximum number of divisions seen in the data. For example, if there were up to 5 divisions found in the data we employed models with 6 time steps. Given that the cultures were incubated for a maximum of 3 days and the mammalian division cycle length of 10-12 hours, it seems highly unlikely that more than 7 divisions occurred in the cultures. However, we checked our results by using up to 9 time steps, allowing for up to 8 divisions, (results not shown in this chapter, but see Chapter 5) and found our results did not change qualitatively. All results contained in this chapter therefore use the minimal number of time steps allowed by the model in order to achieve the number of divisions seen in the data.

4.2.5 Alternative confidence limits for parameter values

In deriving the confidence limits by the Monte-Carlo method (section 3.4) we assume that there is no variance in the data. It seems highly unlikely that repeating the experiments would lead to an identical result as that found in the data. However, in each case we have only one set of results on which to base our limits, albeit that the data consists of a large sample of individual processes. In such a situation we can derive some indication of our the confidence limits given variation in the data by bootstrapping the data (Efron, 1993; Press et al., 2002). This involves resampling the data with replacement with as few assumptions about the data as possible.

Here we assume that the data belongs to a multinomial distribution. The probability of cell belonging to a given division category P_i is therefore given by the fraction of the number of cells in category i with respect to total population of living cells. In mathematical terms this is

$$P_i = \frac{r_i}{\sum_{i=0}^{k} r_i}$$

Where r_i is the observed number of cells in a given division category. The bootstrap procedure is as follows:

1. From the given data set obtain the total number of cells in division categories 0 to k and calculate P_i for $i = 1 \ldots k$.
2. Each cell in the total population of living cells is now randomly assigned to a division category based on the probabilities P_i for $i = 1 \ldots k$.

3. Repeat step 2 a number of times (1000) to create a set of data sets.

4. Estimate the parameters and obtain likelihoods for each data set. For each parameter and likelihood the end result is a distribution of MLEs or likelihoods.

5. Find the bootstrap mean and 95% CIs from these distributions.

4.2.6 Numerical methods for maximizing log-likelihood function

Numerical optimization of the multinomial approximation was achieved using the global optimization method of Stanton et al. (1997). Further cross-checking of our results was done using the Nelder-Mead simplex algorithm (Press et al., 2002).

4.3 Initial Study

4.3.1 Choice of Data

The data found in Hare et al. (1998) are the results of several different experiments, of these, figures, 2, 3, 5 and 6 all contain results of CFSE analysis. In order to produce reliable estimates from our MLE procedure we require the data to contain final cell counts for both DP and SP cells. The results shown in figure 2 pertain to isolated CD4 and CD8 SP cells and are therefore inappropriate for our analysis. Figure 3 of the paper contains data in the form of a time series and this necessarily introduces an added complication that was deemed unsuitable as a starting point for analysis. The data in figure 5 of Hare et al. (1998) are the results of experiments using cells transgenic for the anti-apoptotic protein bcl-2. At the time of our initial analysis we were not sure of what effect this might have on the results. These data were therefore passed over in favour of the data found in figure 6 of Hare et al. (1998).

4.3.2 Figure 6. Hare et al. (1998): The experiment

Work using RTOC preparations carried out by Anderson et al. (1994) suggested that MHC class II^+ epithelial cells were the cell type responsible for the initial stages of positive selection. Further to this work Hare et al. (1998) attempted to assess whether, following positive selection, the subsequent DP to SP transition and SP proliferation was dependent on interaction with the same cells that had initiated positive selection. To this end, these authors used BALB/c and C57BL/6 mice that differed in MHC haplotype: H-2^d and H-2^b respectively. Embryos taken from these mice were

used to prepare both H-2^d and H-2^b thymic stromas. This group typically define stroma consists of cortical epithelium, medullary epithelium and fibroblasts Wilkinson et al. (1995). Subsequently and in separate re-aggregate cultures, positively selected CD69$^+$ DP cells derived from the BALB/c H-2^d strain were added to either the H-2^d or the H-2^b thymic stroma. The idea behind this is that thymocytes that are positively selected on H-2^d MHC will have lower levels of interaction with H-2^b MHC. If thymocyte division occurred in H-2^b cultures then it would seem unlikely that it was entirely due to MHC interaction. The results showed that, as expected, positively selected CD69$^+$ DP cells sourced from H-2^d (matched) mice transited from DP to SP and subsequently divided. However, when DP CD69$^+$thymocytes from H-2^d mice were incubated with mismatched H-2^b stroma they were also seen to behave in a similar fashion. The authors therefore suggest that SP division is due to some cause other than MHC interaction. Subsequently, they suggest that the division enhancing cytokine IL-7 which is expressed by thymic epithelium is the agent responsible.

4.3.3 Result of initial study

For both our models, we used the multinomial approximation, to obtain parameter estimates for the experimental data found in Figure 6. (Hare et al., 1998). We subsequently used these estimates to generate 1000 Monte-Carlo simulated data sets per model. From these data sets we produced a further distributions of MLEs. Comparison of the distribution means with the initial MLEs showed that the estimates produced by the multinomial approximation are unbiased (figure 4.4)[1]. This result was subsequently repeated for all data analysed (Tables 4.4 to 4.9).

The distributions of MLEs also supplied us with 95% CIs for our estimates as described in section 3.4 (Tables 4.4 to 4.9). In addition, based on the assumption that the data is derived from a multinomially distributed population (section 4.2.5) we produced alternative 95% CIs for our estimates (Tables 4.4 to 4.9).

Comparison of the Monte-Carlo simulated data with the experimental data (figure 4.5 and 4.6) suggests that model 2 has a better fit than the model 1 (see table 4.2 for means and 95% CIs). Using our Monte-Carlo likelihood ratio test procedure (section 3.5) this result was shown to be highly significant ($p < .001$). We used Model 1 as the null in this test because Model 2 produced a higher likelihood (see table 4.1)[2]; the aim of the test being to establish whether the best fitting model has a significantly better fit. Graphically, this is shown by the disparity between the log-likelihood ratios obtained from the experimental data and the distribution of the log-likelihood ratios derived from the model 1 simulated data sets (figure 4.7 and table 4.1). This result would suggest that death in these cultures does not occur at the DP stage. Death would therefore appear to occur

[1]Reference to the figures provided by Hare et al. (1998) will take the form of the following example: Figure 6. H-2^d thymic stroma (Hare et al., 1998).

[2]All subsequent tests on the discrete models also used Model 1 as a null.

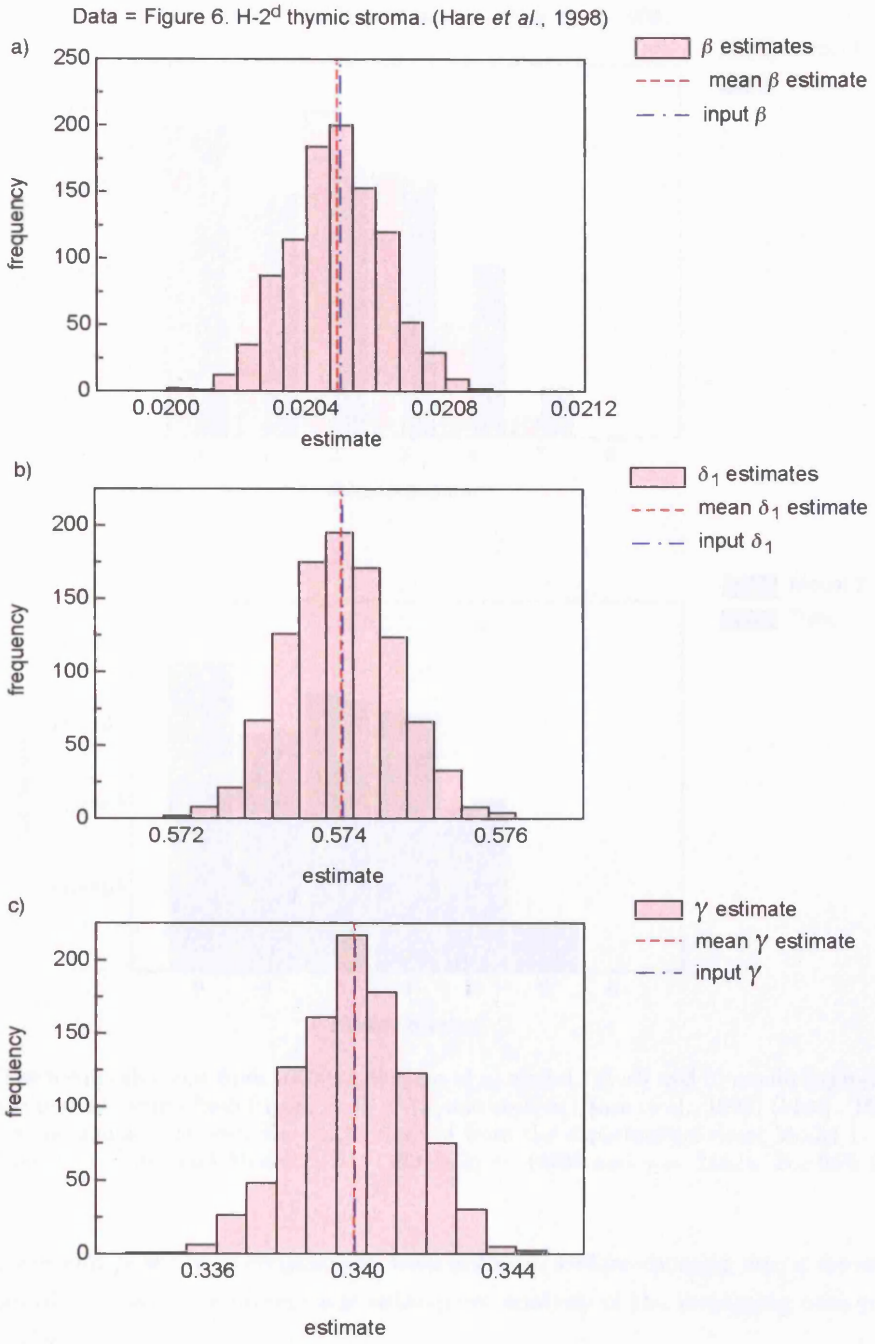

Figure 4.4: The multinomial approximation produces unbiased parameter estimates. Model 1 parameter estimates (input MLEs: $\beta = .02050$, $\delta_1 = .57403$ and $\gamma = .33989$) for the data provided by Figure 6. H-2^d thymic stroma (Hare et al., 1998) were used to create 1000 simulated data sets. The figure shows the distributions of Model 1 MLEs produced by estimating from these data sets. The mean of the distributions and is shown along with the input MLEs used for the simulation. In the case of γ these values are indistinguishable at this scale. See tables 4.4, 4.5 and 4.6 for distribution means and 95% CIs.

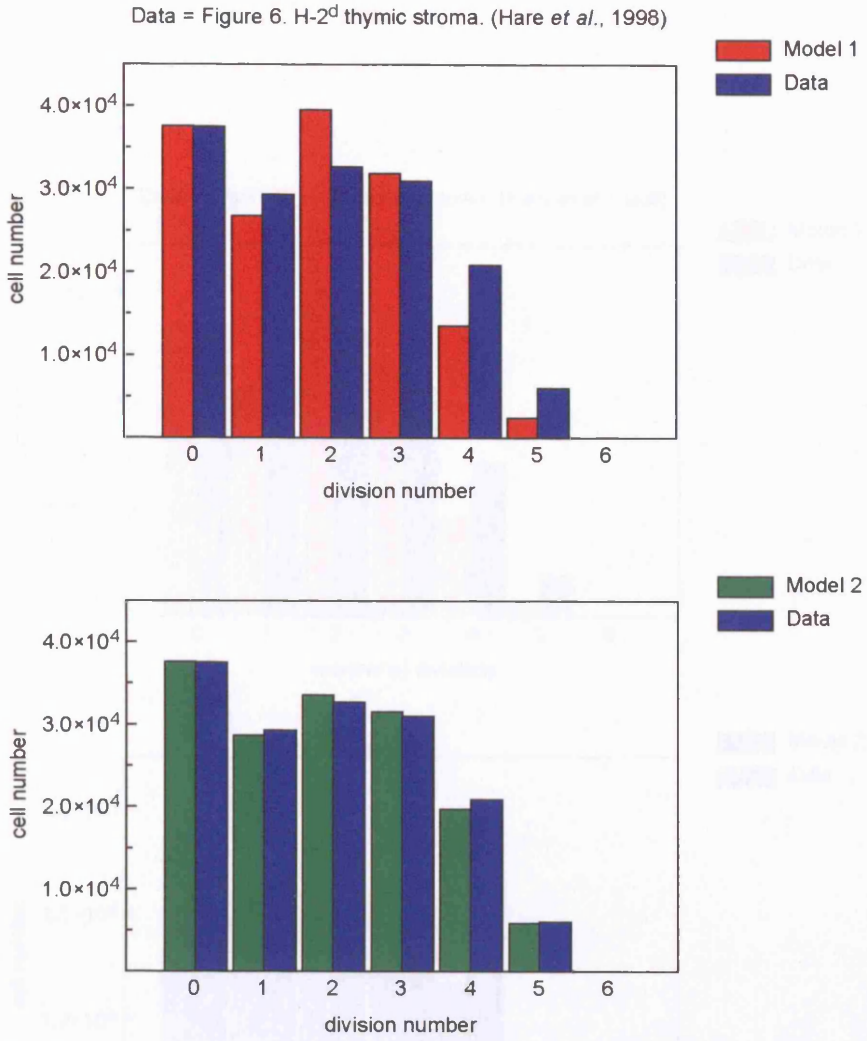

Figure 4.5: The mean cell count from 1000 simulations of a) model 1 (red) and b) model 2 (green) are shown with the experimental results from Figure 6. H-2^d thymic stroma (Hare et al., 1998) (blue). The parameter values used in the simulations were the MLEs derived from the experimental data; Model 1: $\beta = .02050$, $\delta_1 = .57403$ and $\gamma = .33989$ and Model 2: $\beta = .48316$, $\delta_2 = .18789$ and $\gamma = .21624$. For 95% CIs see table 4.2.

at the SP stage and possibly in conjunction with division. Before entering into a discussion of the consequences of this result we present our subsequent analysis of the remaining data published by Hare et al. (1998).

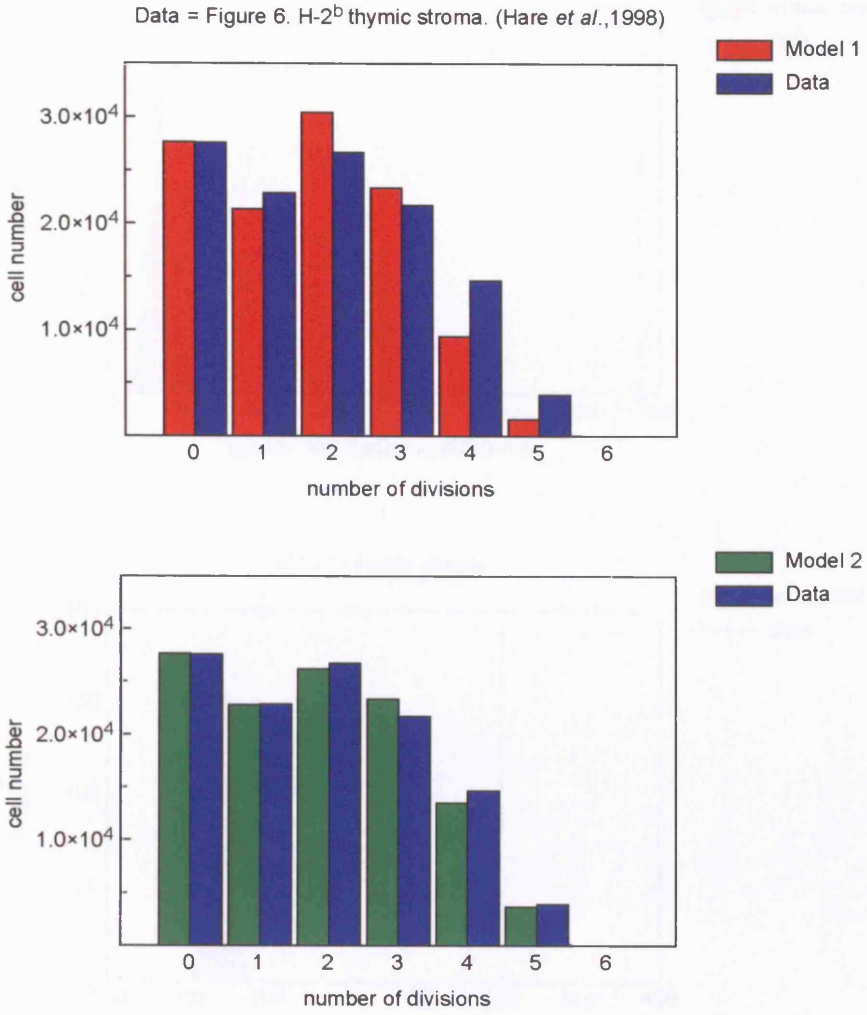

Figure 4.6: The mean cell count from 1000 simulations of a) model 1(red) and b) model 2 (green) are shown with the experimental results from Figure 6. H-2^b thymic stroma (Hare et al., 1998) (blue). The parameter values used in the simulations were the MLEs derived from the experimental data; Model 1: $\beta = .01699$, $\delta_1 = .54286$ and $\gamma = .32362$ and Model 2: $\beta = .51893$, $\delta_2 = .19129$ and $\gamma = .19379$. For 95% CIs see table 4.2.

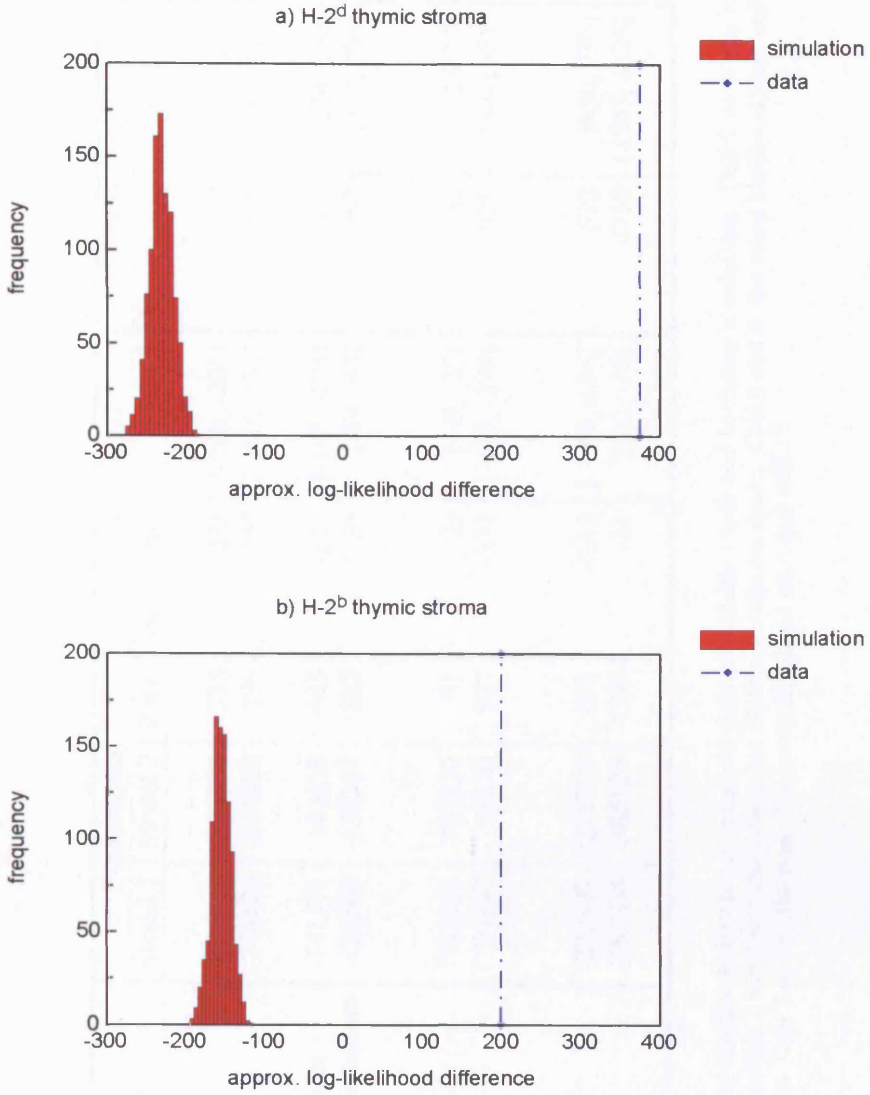

Figure 4.7: Simulated data sets were produced using model 1 MLEs derived from a) Figure 6: H-2^d and b) H-2^b thymic stroma (Hare et al., 1998) as parameter values (see tables 4.4 to 4.9). The figure shows the distributions of log-likelihood ratios (red) derived from these model 1 simulations (1000 per figure). The vertical lines (blue) indicate the respective log-likelihood ratios derived from the experimental data. For 95% CIs see table 4.1.

Source		Experimental			Monte-Carlo		Bootstrap	
Hare *et al.* (1998)		Model 1	Model 2	Ratio	Mean Ratio	95% CI	Mean Ratio	95% CI
Figure 6.								
	H-2[d]	-304733	-304358	375	-231	[-260, -201]	375	[355, 395]
	H-2[b]	-244120	-243920	199	-155	[-180, -131]	199	[182, 215]
Figure 5.								
	Whole Stroma	-141171	-140628	543	-123	[-143, -103]	543	[524, 560]
	Purified Epithelium	-138663	-138381	282	-55	[-69, -42]	282	[272, 293]
Figure 3:								
Start Day 0:								
	Day 2	-781951	-781920	31	-29	[-38, -20]	31	[24, 37]
	Day 3	-754147	-753220	927	-332	[-365, -298]	926	[901, 951]
Figure 3:								
Start Day 1:								
	Day 2	-272397	-271706	691	-0.42	[-1.69, 0.98]	692	[650, 736]
	Day 3	-389834	-385125	4709	-24	[-32, -16]	4710	[4587, 4825]

Table 4.1: The approximate log-likelihoods for the experimental data plus Monte Carlo and bootstrap simulations. Cols 2 to 4: the experimental approx. log-likelihoods under models 1 and 2 and the subsequent likelihood ratio are shown. Cols 5 and 6: the mean likelihood ratio and 95% CI for the Monte Carlo simulated data. Cols 7 and 8: the mean bootstrap likelihood ratio and 95% CI.

Source		Data			Model 1		Model 2	
Hare *et al.* (1998)	Div	Data	Bootstrap	95% CI	MC	95% CI	MC	95% CI
Figure 6. H-2^d:	0	37500	37493	[37142, 37824]	37561	[37206, 37934]	37594	[37236, 37953]
	1	29300	29305	[29016, 29603]	26730	[26323, 27154]	28668	[28266, 29102]
	2	32700	32698	[32390, 33012]	39523	[38924, 40115]	33565	[33062, 34077]
	3	31000	31004	[30692, 31312]	31931	[31280, 32638]	31534	[31020, 32075]
	4	20860	20858	[20578, 21119]	13569	[13078, 14056]	19686	[19274, 20113]
	5	6030	6033	[5881, 6189]	2381	[2200, 2576]	5853	[5602, 6104]
Figure 6. H-2^b:	0	27580	27572	[27287, 27851]	27613	[27286, 27933]	27605	[27299, 27918]
	1	22840	22841	[22570, 23115]	21342	[20982, 21698]	22800	[22430, 23182]
	2	26700	26706	[26415, 26993]	30409	[29825, 30962]	26175	[25757, 26597]
	3	21700	21699	[21440, 21947]	23348	[22743, 23938]	23336	[22880, 23793]
	4	14650	14648	[14426, 14883]	9369	[8979, 9754]	13458	[13117, 13813]
	5	3880	3883	[3761, 4002]	1541	[1396, 1690]	3635	[3454, 3828]
Figure 5. Whole Stroma:	0	13000	12995	[12795, 13200]	13049	[12822, 13272]	13371	[13150, 13582]
	1	13300	13304	[13096, 13509]	11062	[10809, 11330]	11359	[11095, 11638]
	2	12600	12598	[12398, 12815]	18075	[17654, 18492]	14325	[13997, 14641]
	3	14000	14005	[13803, 14216]	15777	[15327, 16220]	15568	[15217, 15935]
	4	10900	10897	[10707, 11096]	7167	[6827, 7508]	12009	[11688, 12353]
	5	9470	9471	[9289, 9654]	1335	[1202, 1478]	4636	[4432, 4866]
Figure 5. Purified Epithelium:	0	10500	10498	[10311, 10689]	10538	[10335, 10736]	10752	[10548, 10949]
	1	14900	14899	[14672, 15126]	12982	[12699, 13263]	13149	[12867, 13424]
	2	15100	15100	[14869, 15319]	20259	[19837, 20696]	17758	[17407, 18120]
	3	16300	16303	[16077, 16516]	16704	[16216, 17153]	17132	[16760, 17512]
	4	11500	11496	[11295, 11687]	7122	[6776, 7475]	10072	[9766, 10365]
	5	5300	5303	[5166, 5432]	1240	[1116, 1376]	2669	[2496, 2834]

Table 4.2: The numerical values for the experimental data found in Figure 6: H-2^d and H-2^b thymic stroma and Figure 5: whole stroma and purified epithelium (Hare et al., 1998) (cols 1 to 3) are shown with their bootstrapped means and 95% CIs (cols 4 and 5). The table also shows the means and 95% CIs derived from Monte-Carlo simulations of model 1 (cols 6 and 7) and model 2 (cols 7 and 8). All means and 95% CIs were taken from 1000 data sets per experimental data source.

Source	Data				Model 1		Model 2	
Hare *et al.* (1998)	Div	Data	Bootstrap	95% CI	MC	95% CI	MC	95% CI
Figure 3:	0	226000	225998	[225519, 226487]	225993	[225140, 226816]	225985	[225122, 226758]
Start Day 0: Day 2	1	61300	61294	[60863, 61716]	60964	[60320, 61613]	61309	[60641, 61928]
	2	24000	24008	[23718, 24292]	25256	[24758, 25763]	23989	[23563, 24425]
	3	5000	5001	[4855, 5133]	3779	[3588, 3962]	5013	[4806, 5222]
Figure 3:	0	151000	150996	[150433, 151588]	151115	[150412, 151803]	151541	[150824, 152293]
Start Day 0: Day 3	1	71400	71397	[70904, 71889]	67232	[66466, 67830]	68611	[67938, 69205]
	2	63300	63307	[62873, 63750]	71663	[70814, 72516]	61509	[60792, 62188]
	3	32600	32594	[32266, 32934]	42367	[41575, 43111]	45437	[44789, 46086]
	4	22400	22404	[22108, 22684]	13264	[12808, 13642]	22856	[22383, 23300]
	5	22400	22402	[22110, 22708]	1726	[1572, 1850]	5567	[5332, 5806]
Figure 3:	0				223743	[223231, 224250]	225958	[225596, 226301]
Start Day 1: Day 2	1				70372	[69631, 70846]	61038	[60466, 61602]
	2				18970	[18539, 19310]	25314	[24861, 25779]
	3				1714	[1576, 1814]	3625	[3442, 3810]
Figure 3:	0				143008	[142455, 143582]	150668	[150231, 151095]
Start Day 1: Day 3	1				99850	[99284, 100441]	67703	[67116, 68270]
	2				71871	[71156, 72603]	69968	[69284, 70647]
	3				25851	[25352, 26351]	43211	[42610, 43874]
	4				4652	[4450, 4845]	14338	[13974, 14730]
	5				335	[280, 392]	1969	[1824, 2124]

Table 4.3: The numerical values for the experimental data found in Figure 3: Days 2 and 3 (Hare et al., 1998) (cols 1 to 3) are shown with their bootstrapped means and 95% CIs (cols 4 and 5). The table also shows the means and 95% CIs derived from Monte-Carlo simulations of model 1 (cols 6 and 7) and model 2 (cols 7 and 8) with initial populations starting at day 0 and day 1. All means and 95% CIs were taken from 1000 data sets per experimental data source.

Source Hare *et al.* (1998)		Model 1: β Approximate Maximum Likelihood Estimates				
		Data	Monte-Carlo	95% CI	Bootstrap	95% CI
Figure 6.						
	H-2[d]	.02050	.02049	[.02023, .02076]	.02050	[.02033, .02067]
	H-2[b]	.01699	.01699	[.01674, .01724]	.01700	[.01683, .01716]
Figure 5.						
	Whole Stroma	.01230	.01230	[.01207, .01254]	.01231	[.01216, .01246]
	Purified Epithelium	.01482	.01482	[.01454, .01509]	.01482	[.01462, .01500]
Figure 3: Start Day 0:						
	Day 2	.03423	.03422	[.03364, .03484]	.03422	[.03378, .03466]
	Day 3	.02501	.02501	[.02476, .02527]	.02501	[.02483, .02520]
Figure 3: Start Day 1:						
	Day 2	0	0	0	0	0
	Day 3	0	0	0	0	0

Table 4.4: Approximate MLEs of β under Model 1 constraints. Col 2: The model 1 estimates derived from the data provided by Hare et al. (1998). Cols 2 and 3: mean estimates and 95% CIs derived from Monte-Carlo simulated data sets (1000 per data source). Cols 3 and 4: mean estimates and 95% CIs taken from bootstrapped data sets (1000 per data source).

Source Hare *et al.* (1998)	Model 1: δ_1 Approximate Maximum Likelihood Estimates				
	Data	Monte-Carlo	95% CI	Bootstrap	95% CI
Figure 6.					
H-2[d]	.57403	.57402	[.57277, .57536]	.57400	[.57281, .57519]
H-2[b]	.54286	.54284	[.54127, .54441]	.54282	[.54138, .54429]
Figure 5.					
Whole Stroma	.49774	.49771	[.49565, .49972]	.49769	[.49590, .49949]
Purified Epithelium	.46736	.46731	[.46470, .46982]	.46733	[.46508, .46976]
Figure 3: Start Day 0:					
Day 2	.59839	.59841	[.59692, .59985]	.59840	[.59736, .59948]
Day 3	.67127	.67125	[.67049, .67209]	.67127	[.67059, .67198]
Figure 3: Start Day 1:					
Day 2	.89588	.89585	[.89502, .89672]	.89587	[.89525, .89650]
Day 3	.85605	.85603	[.85519, .85686]	.85605	[.85523, .85689]

Table 4.5: Approximate MLEs of δ_1 under Model 1 constraints. Col 2: The model 1 estimates derived from the data provided by Hare et al. (1998). Cols 2 and 3: mean estimates and 95% CIs derived from Monte-Carlo simulated data sets (1000 per data source). Cols 3 and 4: mean estimates and 95% CIs taken from bootstrapped data sets (1000 per data source).

Source Hare *et al.* (1998)	Model 1: γ Approximate Maximum Likelihood Estimates				
	Data	Monte-Carlo	95% CI	Bootstrap	95% CI
Figure 6.					
H-2[d]	.33989	.33988	[.33686, .34281]	.33988	[.33800, .34178]
H-2[b]	.32362	.32357	[.32020, .32687]	.32362	[.32153, .32574]
Figure 5.					
Whole Stroma	.34961	.34961	[.34559, .35351]	.34957	[.34640, .35263]
Purified Epithelium	.33192	.33186	[.32818, .33548]	.33193	[.32884, .33503]
Figure 3: Start Day 0:					
Day 2	.21967	.21977	[.21630, .22330]	.21974	[.21729, .22207]
Day 3	.27762	.27759	[.27564, .27960]	.27763	[.27613, .27902]
Figure 3: Start Day 1:					
Day 2	.11872	.11872	[.11775, .11977]	.11872	[.11809, .11935]
Day 3	.15249	.15249	[.15174, .15327]	.15251	[.15210, .15295]

Table 4.6: Approximate MLEs of γ under Model 1 constraints. Col 2: The model 1 estimates derived from the data provided by Hare et al. (1998). Cols 2 and 3: mean estimates and 95% CIs derived from Monte-Carlo simulated data sets (1000 per data source). Cols 3 and 4: mean estimates and 95% CIs taken from bootstrapped data sets (1000 per data source).

Source	Model 2: β Approximate Maximum Likelihood Estimates				
Hare *et al.* (1998)	Data	Monte-Carlo	95% CI	Bootstrap	95% CI
Figure 6.					
H-2[d]	.48316	.48316	[.48166, .48458]	.48320	[.48184, .48456]
H-2[b]	.51893	.51893	[.51721, .52071]	.51898	[.51735, .52057]
Figure 5.					
Whole Stroma	.55868	.55871	[.55680, .56063]	0.55874	[.55673, .56076]
Purified Epithelium	.59745	.59743	[.59499, .60004]	0.59748	[.59485, .5999]
Figure 3.					
Start Day 0:					
Day 2	.47106	.47107	[.46938, .47271]	.47104	[.46985, .47218]
Day 3	.37126	.37125	[.37034, .37216]	.37126	[.37044, .37206]
Figure 3.					
Start Day 1:					
Day 2	.04020	.04017	[.03638, .04403]	.04015	[.03676, .04349]
Day 3	.11076	.11072	[.10937, .11203]	.11077	[.10932, .11217]

Table 4.7: Approximate MLEs of β under Model 2 constraints. Col 2: The model 1 estimates derived from the data provided by Hare et al. (1998). Cols 2 and 3: mean estimates and 95% CIs derived from Monte-Carlo simulated data sets (1000 per data source). Cols 3 and 4: mean estimates and 95% CIs taken from bootstrapped data sets (1000 per data source).

Source Hare *et al.* (1998)	Model 2: γ Approximate Maximum Likelihood Estimates				
	Data	Monte-Carlo	95% CI	Bootstrap	95% CI
Figure 6.					
H-2[d]	.21624	.21623	[.21458, .21778]	.21622	[.21511, .21735]
H-2[b]	.19379	.19379	[.19209, .19548]	.19378	[.19262, .19502]
Figure 5.					
Whole Stroma	.20900	.20897	[.20698, .21088]	.20897	[.20710, .21082]
Purified Epithelium	.18463	.18465	[.18283, .18645]	.18463	[.18303, .18636]
Figure 3.					
Start Day 0:					
Day 2	.10074	.10075	[.09945, .10205]	.10077	[.09989, .10160]
Day 3	.20481	.20483	[.20344, .20623]	.20482	[.20362, .20591]
Figure 3.					
Start Day 1:					
Day 2	.15288	.15289	[.15129, .15446]	.15291	[.15189, .15396]
Day 3	.21724	.21724	[.21604, .21846]	.21725	[.21639, .21811]

Table 4.8: Approximate MLEs of γ under Model 2 constraints. Col 2: The model 1 estimates derived from the data provided by Hare et al. (1998). Cols 2 and 3: mean estimates and 95% CIs derived from Monte-Carlo simulated data sets (1000 per data source). Cols 3 and 4: mean estimates and 95% CIs taken from bootstrapped data sets (1000 per data source).

Source Hare *et al.* (1998)		Model 2: δ_2 Approximate Maximum Likelihood Estimates				
		Data	Monte-Carlo	95% CI	Bootstrap	95% CI
Figure 6.						
	H-2[d]	.18789	.18789	[.18442, .19143]	.18794	[.18529, .19063]
	H-2[b]	.19129	.19127	[.18734, .19519]	.19135	[.18824, .19427]
Figure 5.						
	Whole Stroma	.12848	.12855	[.12399, .13289]	.12858	[.12437, .13268]
	Purified Epithelium	.19860	.19854	[.19460, .20280]	.19861	[.19453, .20242]
Figure 3.						
Start Day 0:						
	Day 2	.14531	.14531	[.13975, .15006]	.14523	[.14156, .14873]
	Day 3	.21074	.21068	[.20746, .21371]	.21073	[.20807, .21363]
Figure 3.						
Start Day 1:						
	Day 2	.69085	.69083	[.68823, .69344]	.69080	[.68872, .69281]
	Day 3	.60040	.60039	[.59860, .60229]	.60038	[.59873, .60204]

Table 4.9: Approximate MLEs of δ_2 under Model 2 constraints. Col 2: The model 1 estimates derived from the data provided by Hare et al. (1998). Cols 2 and 3: mean estimates and 95% CIs derived from Monte-Carlo simulated data sets (1000 per data source). Cols 3 and 4: mean estimates and 95% CIs taken from bootstrapped data sets (1000 per data source).

4.4 Further Results

4.4.1 Results Relating to Fig 5. Hare et al. (1998)

Figure 5. Hare et al. (1998): The experiment

Thymocytes are dependent on maintenance signals from their surrounding stromal environment in order to survive (Anderson et al., 1997, 2000; Anderson and Jenkinson, 2001). In their attempt to assess the role of thymic stromal cells in the developmental events following positive selection Hare et al. (1998) therefore needed to find a method of separating these maintenance signals from signals which affect their development. To this end they used thymocytes derived from bcl-2 transgenic mice that produce DP $CD69^+$thymocytes that express high levels of the anti-apoptotic protein bcl-2. These cells are able to survive in the absence of maintenance signals and therefore can be cultured with or without thymic stroma. A comparison can therefore be made between these 2 regimes and the effects of stroma on development assessed.

The results of this experiment were that when DP $CD69^+$ $bcl\text{-}2^+$ thymocytes are cultured in the absence of thymic stroma they successfully transit to the SP stage but fail to divide. When cultured in the presence of thymic stroma these cells not only transit to the SP stage but also divide. This result indicates that transition to SP stage is independent of stromal support whilst division is not. Furthermore, when purified epithelium replaced whole stroma in these cultures division is also seen to occur. This may indicate that thymic epithelium is responsible for inducing the proliferative response.

Figure 5. Hare et al. (1998): Results

Once again we proceeded by using the multinomial approximation to estimate parameter values for both models from the data. These MLEs were then used to simulate 1000 data sets per model. Subsequently, distributions of MLEs were produced from these data sets from which we obtained distribution means and 95% CIs (tables 4.4 to 4.9). As previously noted these results show that the multinomial approximation is unbiased.

The results of the comparison between the means of the simulated data with the experimental data (figure 4.8, 4.9 and see table 4.2 for numerical values) are less convincingly in favour of hypothesis 2 (model 2) than those for the initial study. However, there is a large disparity between the log-likelihood ratios obtained from the experimental data and the distribution of the log-likelihood ratios derived from model 1 simulated data sets (figure 4.10 and table 4.1). The result is therefore

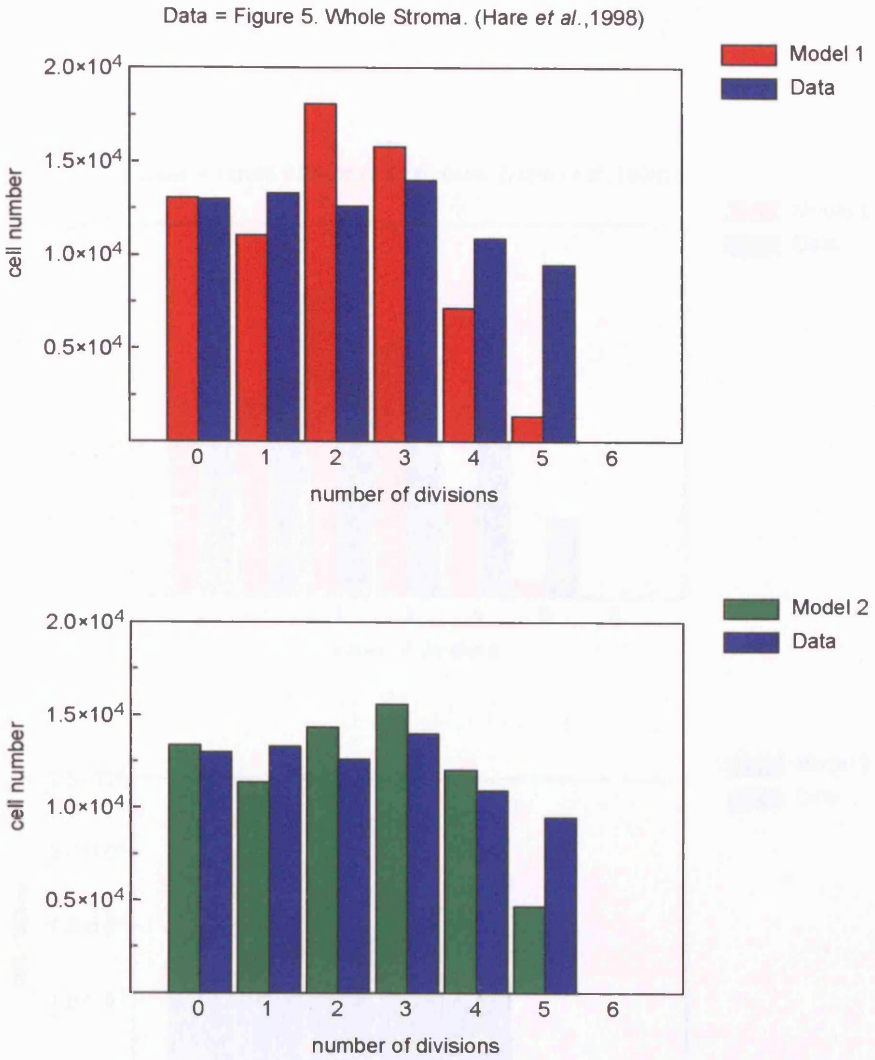

Figure 4.8: The mean cell count from 1000 simulations of a) model 1 (red) and b) model 2 (green) are shown with the experimental results from Figure 5. Whole thymic stroma (Hare et al., 1998) (blue). The parameter values used in the simulations were the MLEs derived from the experimental data; Model 1: $\beta = .01230$, $\delta_1 = .49774$ and $\gamma = .34961$ and Model 2: $\beta = .55868$, $\delta_2 = .12848$ and $\gamma = .20900$. For 95% CIs see table 4.2.

that model 2 has a better fit than model 1 ($p < .001$). This result suggests that death in these cultures does not occur at the DP stage. Consequently, this result also suggests that death occurs at the SP stage and it could occur in conjunction with division. More generally, this result is similar to the result of our initial study and therefore suggests that regardless of the stromal types or thymocytes used in the experiments the underlying pattern of behaviour is consistent.

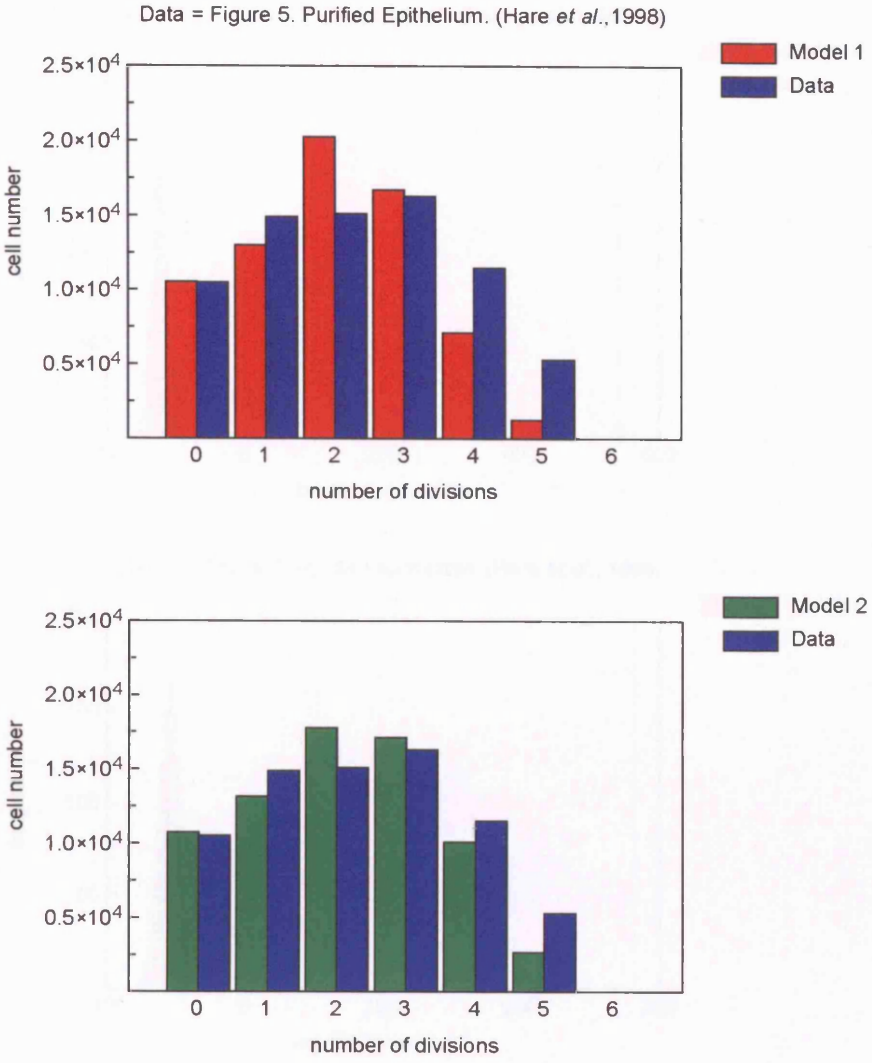

Figure 4.9: The mean cell count from 1000 simulations of a) model 1 (red) and b) model 2 (green) are shown with the experimental results from Figure 5. Purified epithelium (Hare et al., 1998) (blue). The parameter values used in the simulations were the MLEs derived from the experimental data; Model 1: $\beta = .01482$, $\delta_1 = .46736$ and $\gamma = .33192$ and Model 2: $\beta = .59745$, $\delta_2 = .19860$ and $\gamma = .18463$. For 95% CIs see table 4.2.

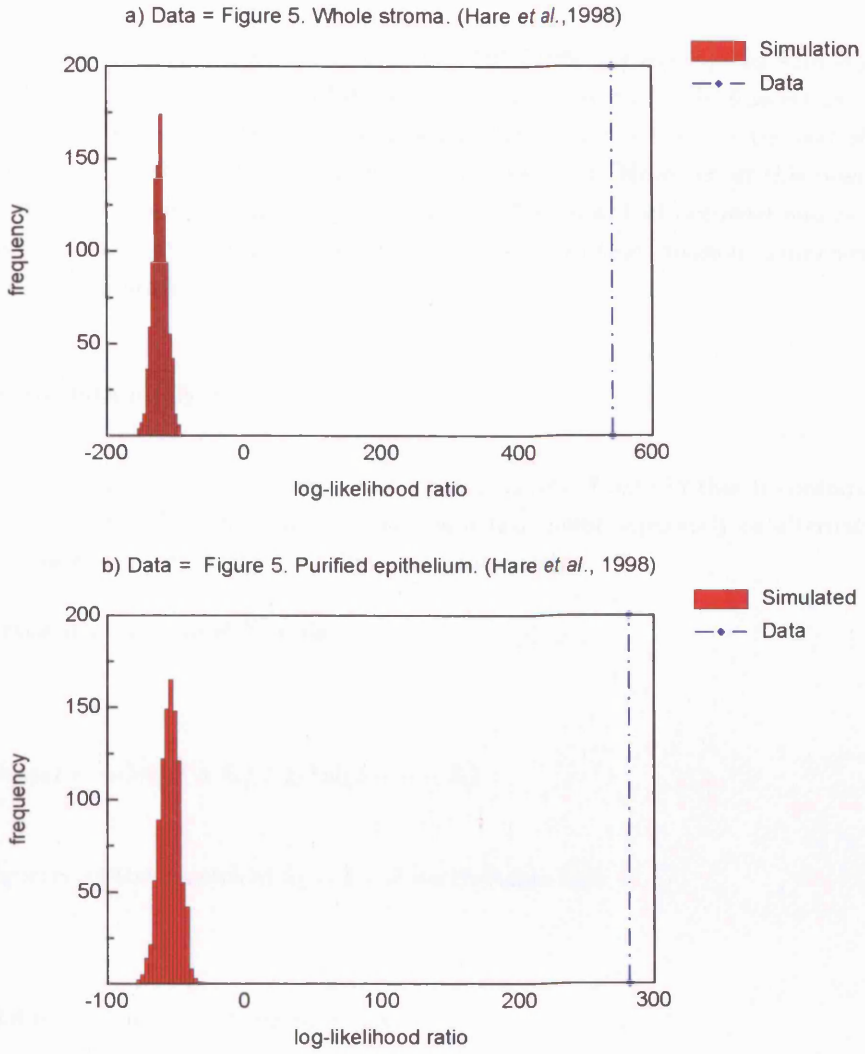

Figure 4.10: Simulated data sets were produced using model 1 MLEs derived from a) Figure 5: Whole thymic stroma and b) purified epithelium (Hare et al., 1998) as parameter values (see tables 4.4 to 4.9). The figure shows the distributions of log-likelihood ratios (red) derived from these model 1 simulations (1000 per figure). The vertical lines (blue) indicate the respective log-likelihood ratios derived from the experimental data. For 95% CIs see table 4.1.

4.4.2 Results Relating to Fig 3. Hare et al. (1998)

Figure 3. Hare et al. (1998): The experiment

Figure 3. (Hare et al., 1998) is a time series. Here DP CD69$^+$, re-aggregated with thymic stroma, were incubated for periods of 1, 2 and 3 days and then analysed for CFSE fluorescence. In addition the expression of CD4 and CD8 was also observed. The result is that, at the end of day 1 some DP cells have successfully made the transition to the SP stage. However, at this point no division appears to have occurred. By day 2 up to 3 rounds of division had occurred and by day 3 up to 5 had occurred. Hare et al. (1998) state that this indicates that division commences only after transition to the SP stage.

Approach to data analysis

Figure 3. (Hare et al., 1998) differs from our 2 previous sets of data in that it contains data in the form of a time series. We can choose to treat each time point separately or alternatively we can combine the data. First we treat each time point separately.

After one time step the model 2 yields

$$\ell(\theta|data_1) = d_0 \log(\beta + \delta_1) + \xi_l \log(1 - \beta - \delta_1) \tag{4.10}$$

as a consequence of the constraint $\delta_1 = 1 - \beta$ we therefore find

$$\ell(\theta|data_1) = d_0 \log(1) + \xi_l \log(0) = -\infty \tag{4.11}$$

Thus after one time step model 2 does not fit. This might be construed as evidence that model 2 is a poor model. However, we argue that the lack of ability of the model to provide a usable estimator for this time point is an artifact of the discrete nature of the model. The consequence of this artifact is to place a constraint on the model whereby it is only valid at time steps greater than 1. Furthermore, this constraint not only mitigates the separate analysis of the day 1 data set but also has further implications relating to its inclusion in any combinations of data sets (see below).

However, the data provided for day 1 can be used as a starting population, as an accompanying flow cytometry plot gives the proportions of DP and SP cells at this point in time. We refer to these proportions as η and ν respectively. In this situation we derive the expectation of the multi-type branching process $E(Z_k|Z_0)$ (see equation 2.8) by setting the first entry in the k dimensional vector Z_0 to be η and the second ν, whilst the remaining entries remain as zeros. All subsequent steps of our derivation of the estimating function remain as above.

Given the above we therefore provide analysis based on:

1. Data from day 3 alone
2. Data from day 2 alone

for initial time points: day 0 and day 1.

Results based on day 3 alone

Here we followed the same procedure that we used for both the initial study and the analysis of the data from Figure 5. (Hare et al., 1998). Once again our data produce evidence that the estimates obtained through the multinomial approximation are unbiased (see tables 4.4 to 4.9).

The result of the comparison between the means of simulated data sets with the experimental data given initial conditions are set at day 0 is ambiguous in that it is not clear which model provides the better fit (figure 4.11 and table 4.3). When initial conditions are set at day 1 (figure 4.12) the result suggest that model 2 provides a better fit to the data than model 1. However, in both cases, the large disparity between the log-likelihood ratios obtained from the experimental data and the distribution of the log-likelihood ratios derived from the model 1 simulated data sets is unequivocal (figure 4.13). In both cases, therefore, the model 2 has a significantly better fit than the model 1 ($p < .001$).

This result suggests that death in these cultures does not occur at the DP stage. In addition, the result also suggests that death occurs at the SP stage and could occur in conjunction with division. This result is consistent with those obtained our previous analyses.

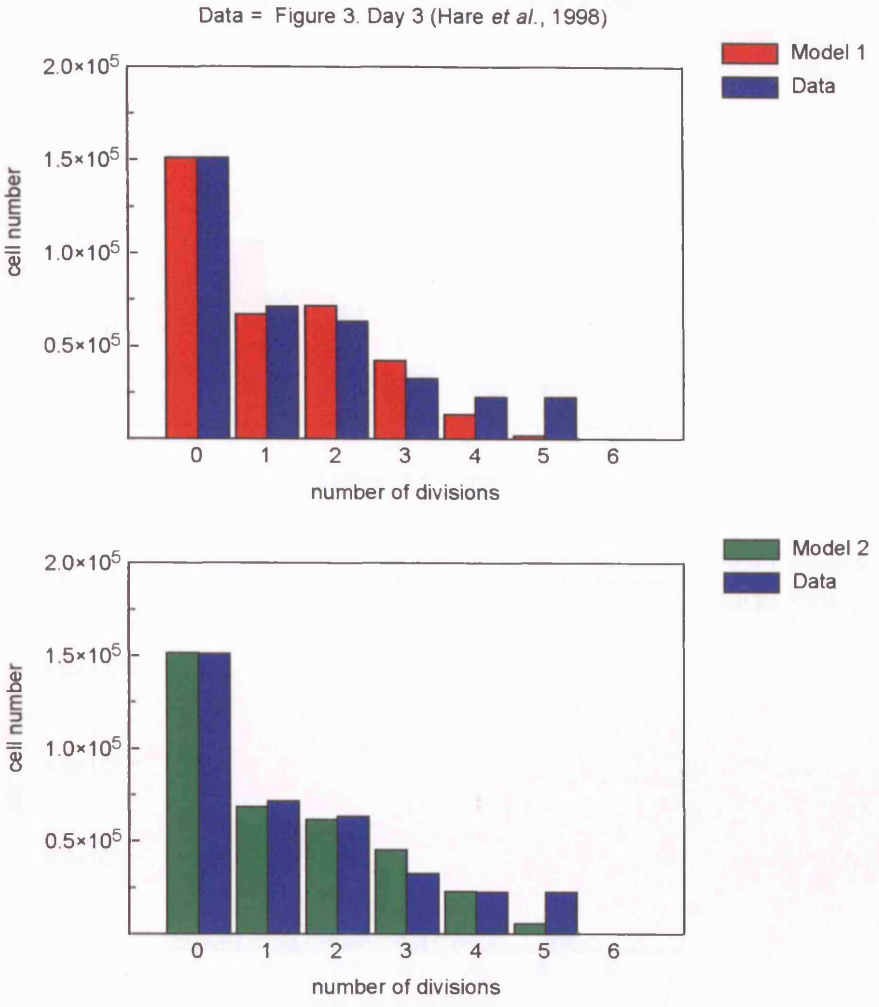

Figure 4.11: The mean cell count from 1000 simulations of a) model 1 (red) and b) model 2 (green) are shown with the experimental results from Figure 3. Day 3 (Hare et al., 1998) (blue). The parameter values used in the simulations were the MLEs derived from the experimental data with initial conditions as at Day 0; Model 1: $\beta = .02501$, $\delta_1 = .67127$ and $\gamma = .27762$ and Model 2: $\beta = .37126$, $\delta_2 = .21074$ and $\gamma = .20481$. For 95% CIs see table 4.3.

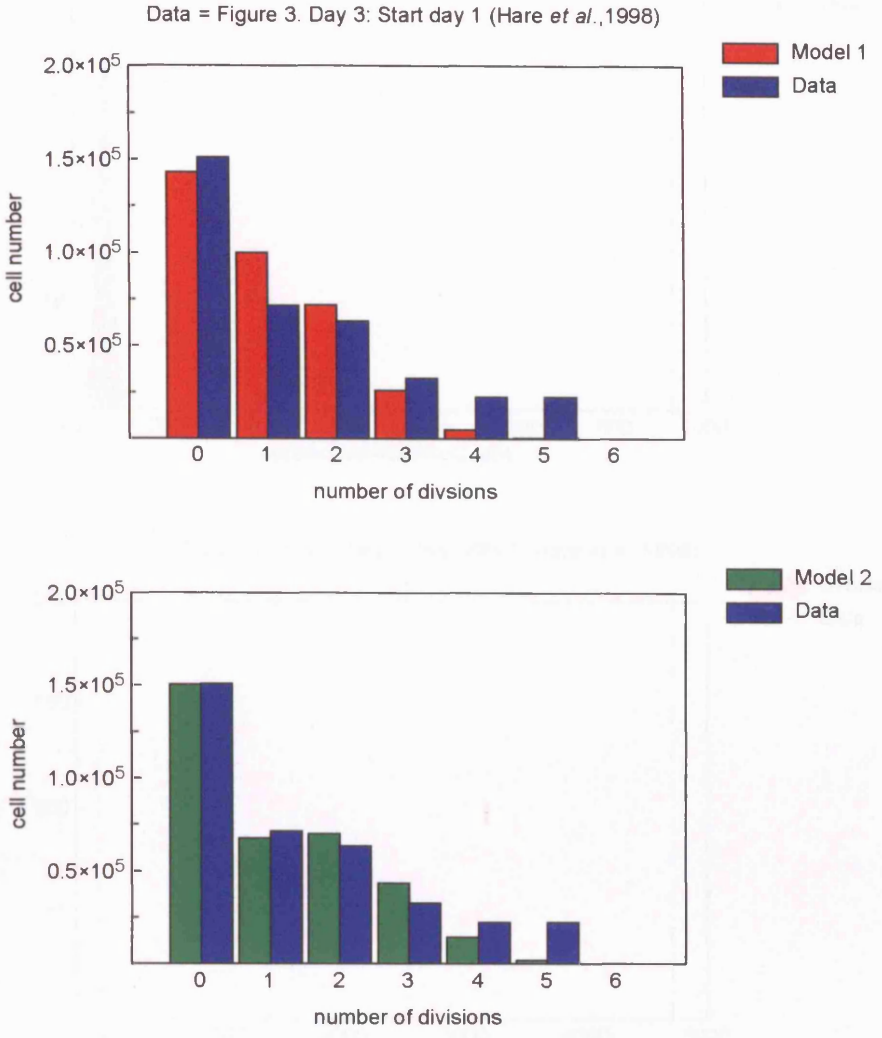

Figure 4.12: The mean cell count from 1000 simulations of a) model 1 (red) and b) model 2 (green) are shown with the experimental results from Figure 3. Day 3 (Hare et al., 1998) (blue). The parameter values used in the simulations were the MLEs derived from the experimental data with initial conditions as at Day 1; Model 1: $\beta = 0$, $\delta_1 = .85605$ and $\gamma = .15249$ and Model 2: $\beta = .11076$, $\delta_2 = .60040$ and $\gamma = .21724$. For 95% CIs see table 4.3.

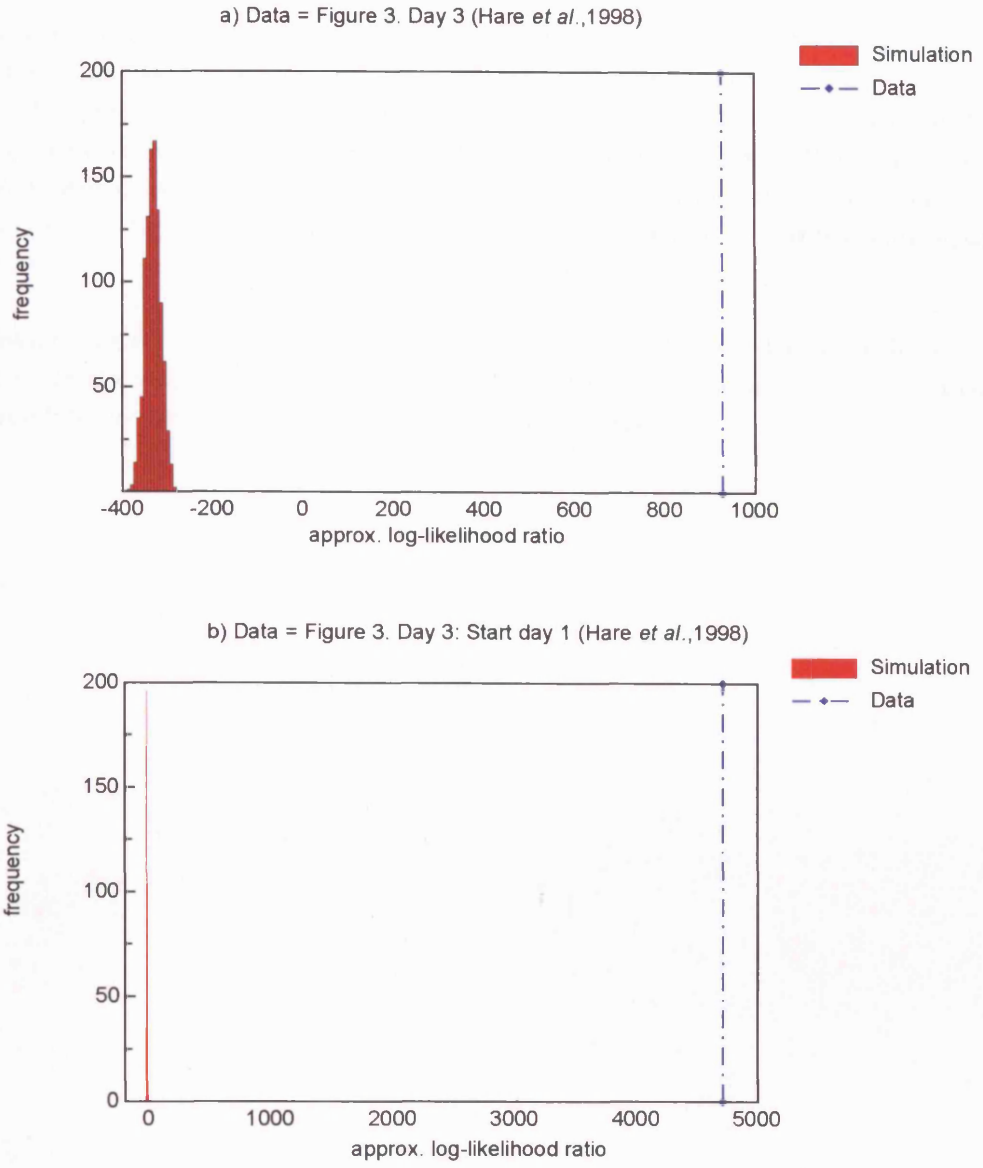

Figure 4.13: Simulated data sets were produced using model 1 MLEs derived from Figure 3: Day 3 (Hare et al., 1998) as parameter values (see tables 4.4 to 4.9). The figure shows the distributions of log-likelihood ratios (red) derived from these model 1 data sets (1000 per figure) given two different initial conditions: the initial population set to the experimentally observed value at a) Day 0 and b) Day 1. The vertical lines (blue) indicate the respective log-likelihood ratios derived from the experimental data. For 95% CIs see table 4.1.

Results based on day 2 alone

Repetition of our analysis procedure also produced more evidence that the multinomial approximation is an unbiased estimator (see tables 4.4 to 4.9). Comparison of the means of the simulated data with experimental data for day 2 alone, given initial conditions set to day 0, suggest that there is little difference between the fit of the models (figure 4.14 and table 4.3). When initial conditions are set to day 1, the fit is also similar with model 2 appearing to provide the slightly better fit (figure 4.15). In both cases, when subjected to our significance test we find that once again model 2 has a better fit than model 1 ($p < .001$) (figure 4.16).

This result suggests that death in these cultures does not occur at the DP stage. In addition, the result would also suggest that death therefore occurs at the SP stage in conjunction with division. This result is also consistent with all of our previous analysis.

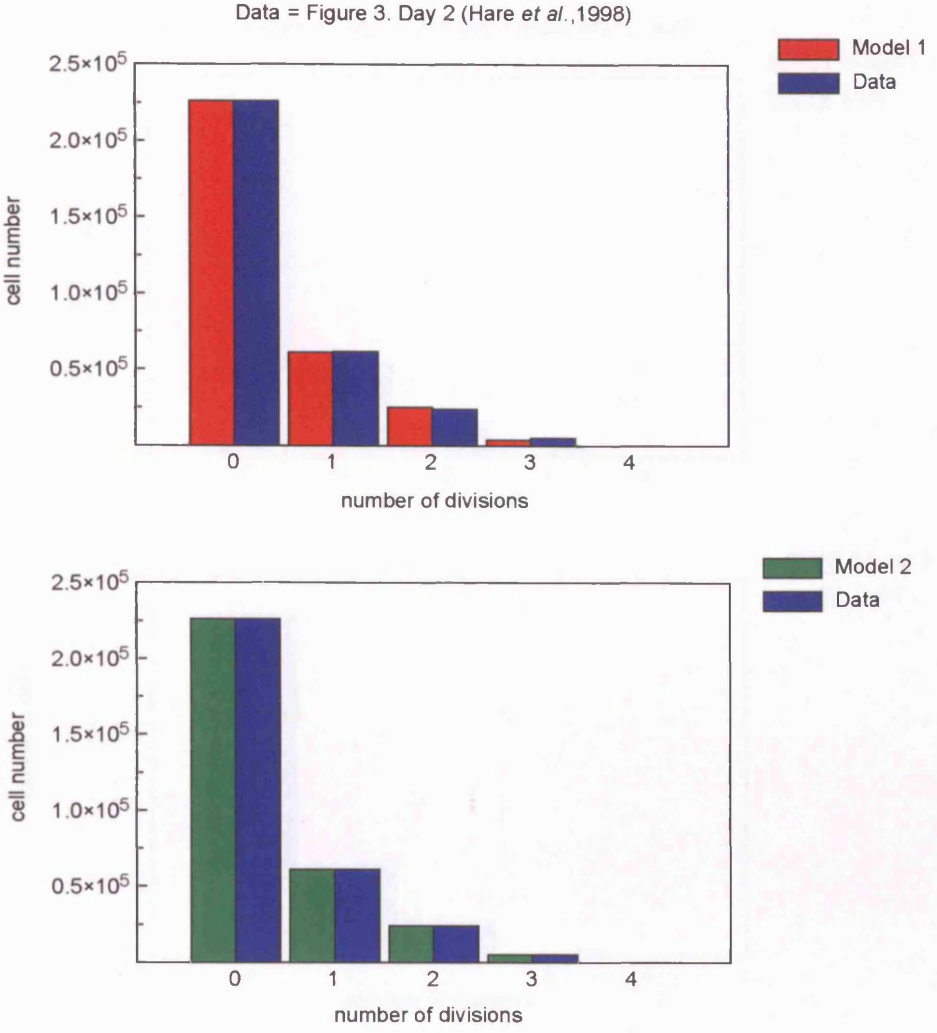

Figure 4.14: The mean cell count from 1000 simulations of a) model 1 (red) and b) model 2 (green) are shown with the experimental results from Figure 3. Day 2 (Hare et al., 1998) (blue). The parameter values used in the simulations were the MLEs derived from the experimental data with initial population as at Day 0; Model 1: $\beta = .03423$, $\delta_1 = .59839$ and $\gamma = .21967$ and Model 2: $\beta = .47106$, $\delta_2 = .14531$ and $\gamma = .10074$. For 95% CIs see table 4.3.

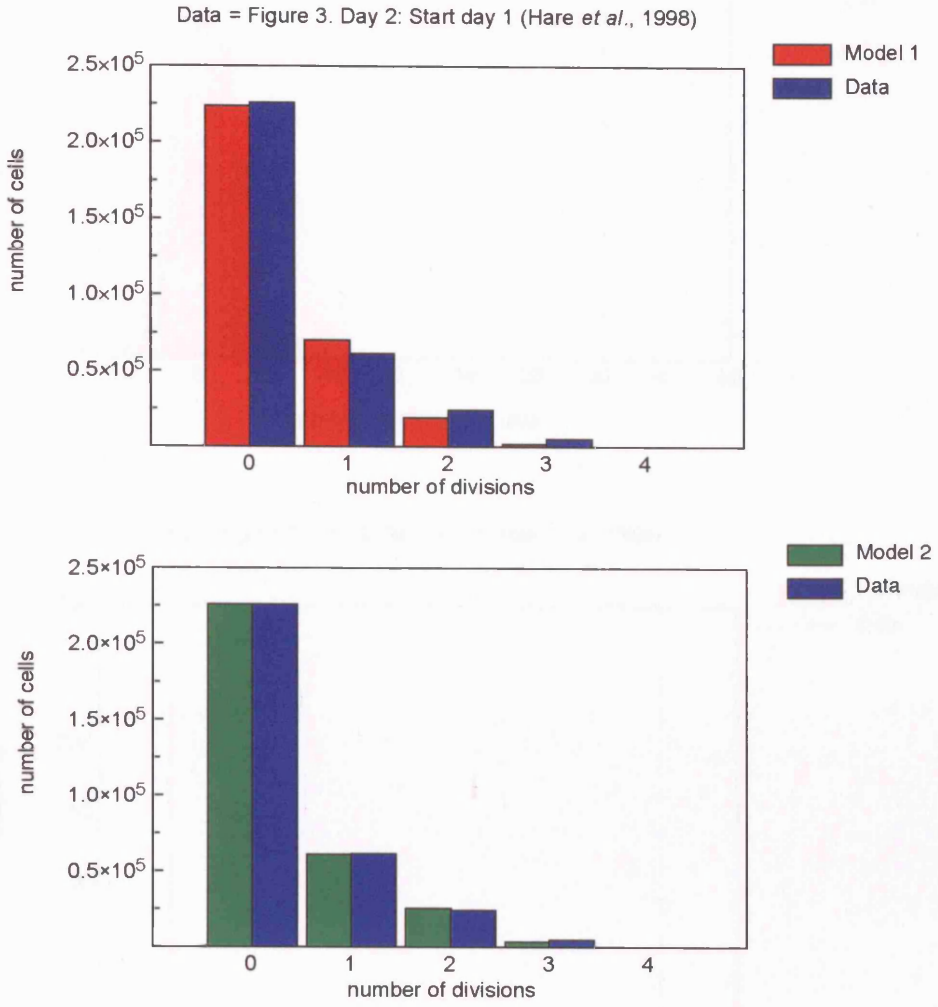

Figure 4.15: The mean cell count from 1000 simulations of a) model 1 (red) and b) model 2 (green) are shown with the experimental results from Figure 3. Day 2 (Hare et al., 1998) (blue). The parameter values used in the simulations were the MLEs derived from the experimental data with initial conditions as at Day 1; Model 1: $\beta = 0$, $\delta_1 = .89588$ and $\gamma = .11872$ and Model 2: $\beta = .04020$, $\delta_2 = .69085$ and $\gamma = .15288$. For 95% CIs see table 4.3.

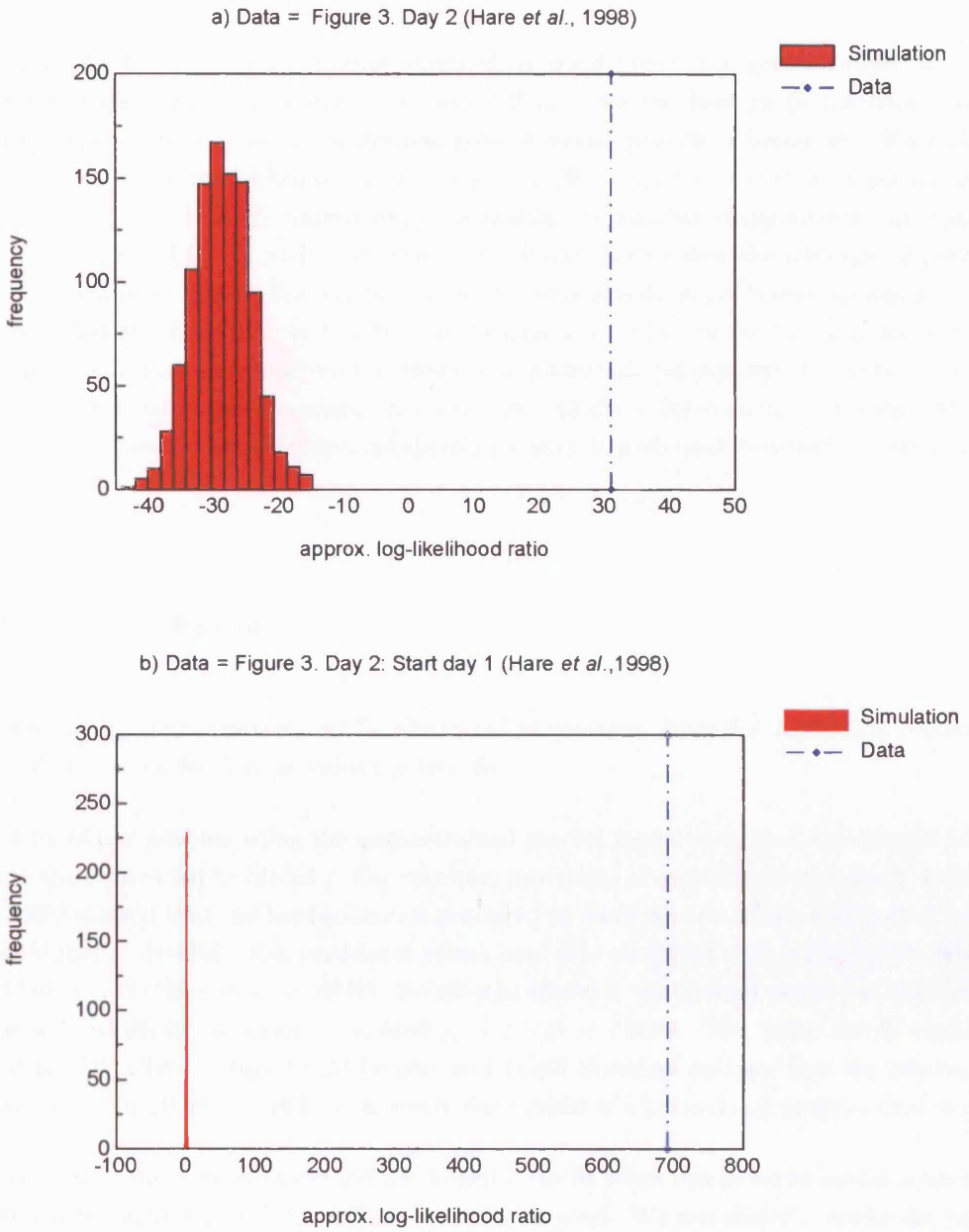

Figure 4.16: Simulated data sets were produced using model 1 MLEs derived from Figure 3: Day 2 (Hare et al., 1998) as parameter values (see tables 4.4 to 4.9). The figure shows the distributions of log-likelihood ratios (red) derived from these model 1 data sets (1000 per figure) given two different initial conditions: the initial population set to the experimentally observed value at a) Day 0 and b) Day 1. The vertical lines (blue) indicate the respective log-likelihood ratios derived from the experimental data. For 95% CIs see table 4.1.

4.4.3 Fitting the General Model

Thus far we have compared the results obtained by constraining the general model to produce 2 distinct models. Our results show that model 2 provides the best fit to the data. Next we asked the question would the unconstrained general model provide a better fit? Note that the unconstrained general model has one more parameter ($\boldsymbol{\theta} = \{\delta_1, \beta, \delta_2, \gamma\}$) than either models 1 or 2 ($\boldsymbol{\theta} = \{\delta_1, \beta, \gamma\}$ or $\{\beta, \delta_2, \gamma\}$ respectively). Increasing the number of parameters can sometimes result in an improved fit. In such cases the usual practice is to follow the principle of parsimony, whereby the simplest explanation (or model) of the data is used in preference to a more complex model provided the difference in the fit is not significant. This raises the question of whether any improvement is significant given the presence of additional parameters. A number of methods have sought to resolve this question. For example, Akaike's Information Criterion (AIC) is a frequently used method that involves maximizing a models likelihood weighted by the number of its parameters:

$$\text{AIC} = 2h - 2\ell(\boldsymbol{\theta} \mid data)$$

where h is the number of parameters for the model in question. Here the model that produces the lowest AIC is the model that provides the best fit.

The results of our analysis using the unconstrained general model were that it produced identical results to those provided by Model 2. For example, the results of our analysis of figure 6. H-2^d (Hare et al., 1998) showed that the log-likelihoods produced by both models where identical: Model 2 = General Model = -304358. The parameter values here also coincided with both models producing $\beta = .48316$, $\gamma = .21624$ and $\delta_2 = .18789$. Relative to Model 2, the general model has the additional parameter δ_1 which in our example yielded $\delta_1 = 1 - \beta = .58684$. This latter result would mean death at the DP $CD69^+$ stage would be zero and would therefore indicate that the general model yields the same result as Model 2. This result was typical of all remaining analysis (not shown).

Considering that the general model did not improve the fit when compared to model 2, there is no need to test for significance using AIC or a similar method. We can directly invoke the principle of parsimony and state that, in comparison to the general model, Model 2 provides the best fit to the data.

4.5 Combining data from different days

Given a specified model, we can use more than one data set in determining the likelihood. Since given data sets numbered 1 to j:

$$\ell(\theta \mid dat_1, dat_2, \ldots, dat_j) = \ell(\theta \mid dat_1) + \ell(\theta \mid dat_2) + \ldots + \ell(\theta \mid dat_j) \tag{4.12}$$

Figure 3 (Hare et al., 1998) contains 3 separate data sets and these are all derived from the identical experimental initial conditions. Equation 4.12 therefore suggests that we can combine these data sets as yet another analysis of the data. However, above we saw that data from day 1 is unsuitable for analysis as a separate data set. The difficulty caused by the inabiltiy of model 2 to fit the data at day 1 also means that it must be excluded from use in combination with others.

In addition to this issue we must also consider another possible difficulty in combining data sets. The destructive nature of the experimental sampling procedure means that each days data was provided by separate cultures. If for unknown experimental reasons growth/death rates in these individual cultures differ widely we would obviously find that our parameter estimates would also differ widely. In such a case it would be pointless to combine data from different days because the data would belong to different statistical populations. Excluding data from day 1 for the reasons above, the question is therefore: for a given model do the data for day 2 and day 3 belong to the same statistical population. We answer this question by constructing a further significance test based on our Monte-Carlo procedures: the population test.

4.5.1 The population test: a test for comparison of data from different days

In order to answer this question we see that the equation 4.12 enables us to construct a further Monte-Carlo test of significance, since for n sets of data the "deviance" $D(\ell)$ is given by

$$D(\ell) = \ell(\theta \mid dat_1) + \ell(\theta \mid dat_2) + \ldots + \ell(\theta \mid dat_j) - \ell(\theta \mid dat_1, dat_2, \ldots, dat_j) \tag{4.13}$$

If the data on two different days belong to the same statistical population then $D(\ell) = 0$.

For a given model we therefore take the following steps

1. Estimate the likelihoods for the experimental data of day 2 alone, day 3 alone and days 2 and 3 combined.

2. Use the parameter estimates obtained from either one of day 2 or day 3 and generate a large number (1000) of data sets for both days 2 and day 3.

3. Estimate the likelihoods for each data set as in step 1.

4. Use equation 4.13 and each triplet of likelihoods generated by steps 2 and 3 to obtain a distribution of $D(\ell)$.

5. Compare this distribution to the value of experimentally derived value of $D(l)$; use the likelihoods obtained from the experimental results (step 1.).

6. Use the previously described method to obtain a value for p.

The results of the population test revealed that for both our models data at day 2 differs significantly from that of day 3 (figure 4.17). This was true regardless of the start day (result starting at day 0 not shown). This result suggested that we could not combine data from day 2 with day 3 and could only examine the observations from different days separately.

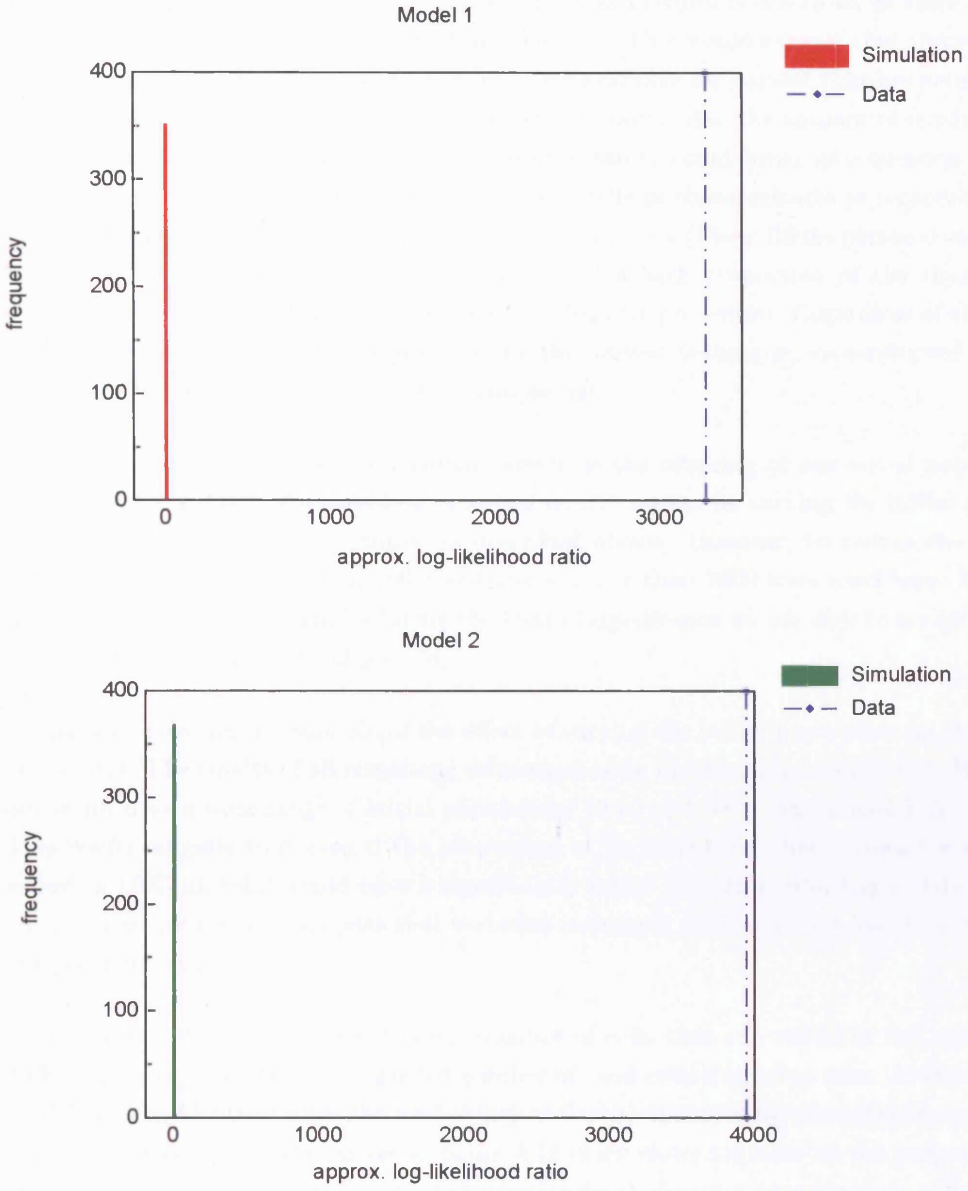

Figure 4.17: The distribution of $D(\ell)$ (equation 4.13) for Model 1 (red) and Model 2 (green) obtained from simulated data sets (1000 per figure) as described in section 4.5.1. The initial population was that observed at Day 1. Results using initial population as at Day 0 were similar (not shown). The vertical line (blue) is the experimentally derived value of $D(\ell)$.

4.6 Robustness of results to sample size

The experimental method for harvesting cells in these cultures is described by Hare et al. (1998) as a process of "gently teasing...apart with fine knives". This would suggest that there is a great deal of uncertainty as to the proportion of thymocytes available for harvest that are actually harvested. Moreover, our results are dependent on the data. In particular, the amount of dye lost is indirectly inferred from the numbers of live cells. The uncertainty could bring into question the legitimacy of our results. However, the final harvest of the cells in these cultures is regarded as a valuable commodity since their production is difficult and laborious (Mary Ritter personal communication). It thus seems likely that such harvests may yield a high proportion of the thymocytes in the culture despite the possibly crude nature of the harvest procedure. Regardless of this and in view of the uncertainty in sample size provided by the harvest technique, we conducted a check on the robustness of our results to variation in sample size.

Given that the final sample is a random sample of the offspring of our initial population we can easily assess the effect of incomplete sampling on our results by varying the initial population and repeating our statistical procedures as described above. However, to reduce the time taken to obtain our results, data sets of 100 simulations rather than 1000 were used here. This makes our significance test less powerful reducing the level of significance we are able to accept to a lower but biologically significant level of $p < .01$.

Figure 4.18 provides an example of the effect of varying the initial population on the robustness of our results. The results of all remaining robustness tests were similar (not shown). Here we see that our result over a wide range of initial populations (0.10 to 1.50 $\times$ experimental initial population). This result suggests that even if the proportion of thymocytes gathered from those available was as low as 10%, model 2 would have a significantly better fit than model 1 ($p < .01$). This supports our previous results and suggests that variation in sample size would not lead to a contradiction of our previous results.

Note however that there is a minimum number of cells that can reside in the initial population. This minima occurs when our expected number of dead cells is equal to zero. At this point both our models will be identical since the probability of death, whatever the phenotypic stage, is also zero. This is the reason that the curves in figure 4.18 draw closer together as the proportion decreases. However, it seems highly unlikely that thymocyte death does not occur in these cultures. Given the value of thymocytes produced by these cultures (see personal communication Mary Ritter above) we therefore suggest that this possibility be ignored.

In addition to the above we also include the results as they relate to our parameter estimates derived from the data data provided by Figure 6: H-2^d (Hare et al., 1998) (figure 4.19). In the

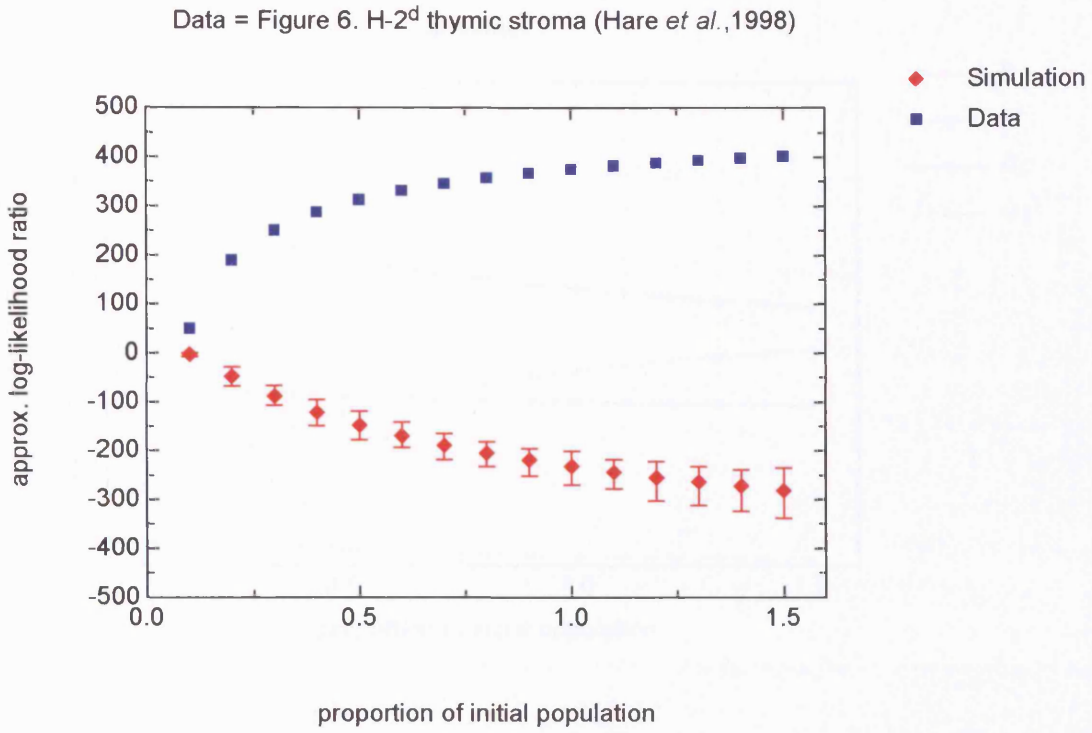

Figure 4.18: Typical result of robustness study. The figure shows the Monte Carlo significance test results when the initial population is varied (shown as a proportion of the intial population). The experimental data was sourced from Figure 6: H-2^d thymic stroma (Hare et al., 1998). The initial experimental population was 8×10^5. The error bars show the upper and lower limits of the simulated approximate likelihood distribution.

case of model 1, the figure shows that providing the proportion of thymocytes harvested is greater than around 50% the principal parameter values (β, γ, and δ_1) are relatively insensitive to sample size. However, the probability of death ($\alpha_1 = 1 - \delta_1 - \beta$) is less robust to changes in sample size. Given that the gradients of the curves for both β and δ_1 share the same sign, this later result is perhaps unsurprising. In the case of Model 2 parameters (β, γ, and δ_2), we find that they are more sensitive to sample size. Once again we see that the probability of death ($\alpha_2 = 1 - \delta_2 - \gamma$) is the less robust to changes in sample size. In a similar vein to model 1 this is a probable reflection of the common gradient sign of γ and δ_2. We reserve comment on the implications of the robustness of parameter estimates to sample size until our discussion (section 4.7).

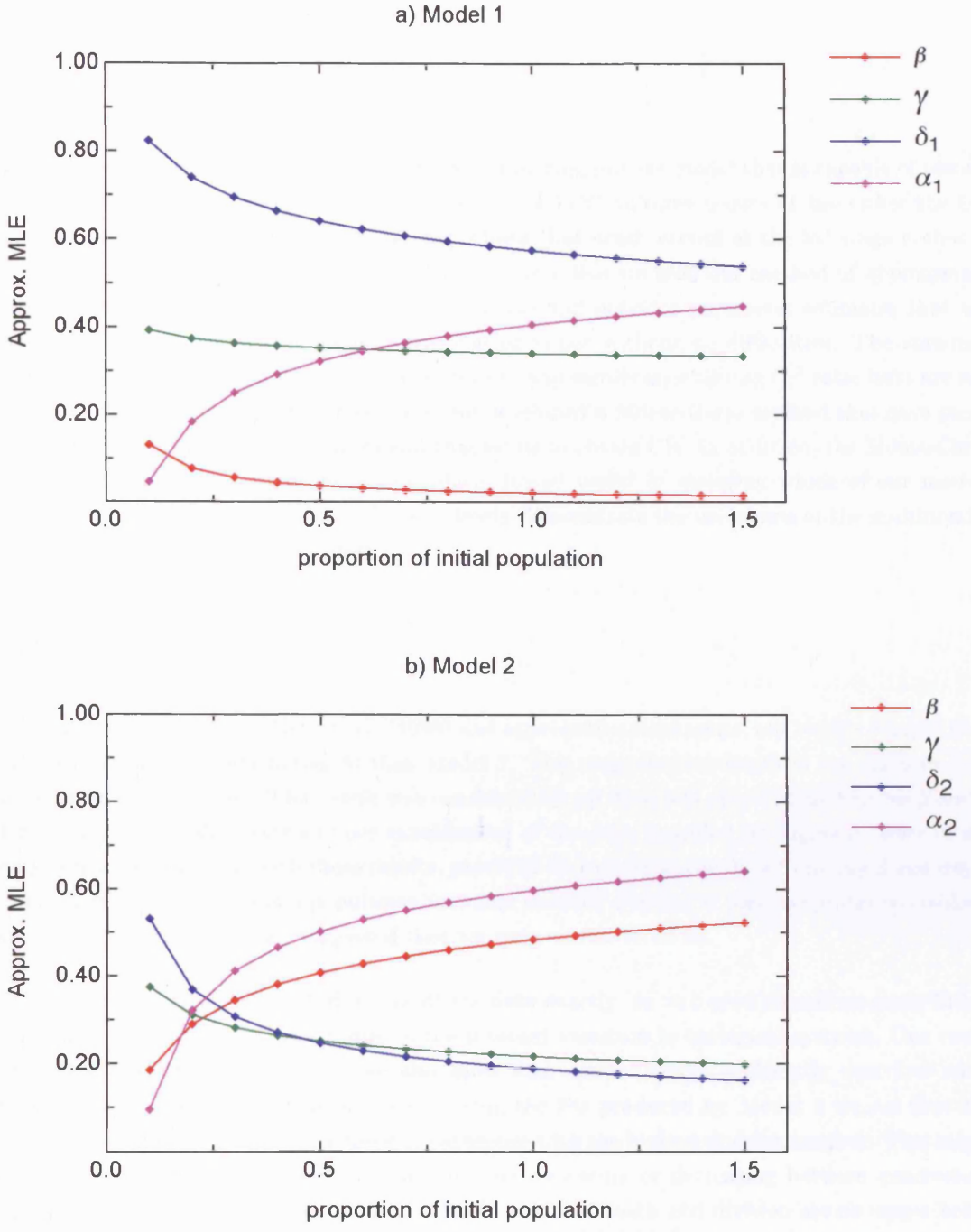

Figure 4.19: The effect of varying the initial population on approximate MLEs. The figure shows how MLEs derived from the data provided by Figure 6: H-2^d thymic stroma (Hare et al., 1998) using a) Model 1 and b) Model 2 alter when the initial population is varied. The initial experimental population was 8×10^5 thymocytes.

4.7 Discussion

4.7.1 The model

Here we have constructed a general discrete time branching process model that is capable of testing our specific hypotheses; that death in the examined FTOC cultures occurs at the either the DP stage or the SP stage. An adjunct to the hypothesis that death occurs at the SP stage is that it occurs in conjunction with division. In order to achieve this we used our method of approximate maximum likelihood. The results show that this method provides parameter estimates that are unbiased. However, the multinomial approximation is not without its difficulties. The standard methods of obtaining CIs (quadratic approximation) and significance testing (χ^2 ratio test) are not applicable. In order to overcome these issues we developed a Monte-Carlo method that gave proof of the unbiased nature of our estimates and enabled us to obtain CIs. In addition, the Monte-Carlo method also produced a significance test which proved useful in assessing which of our models gave the better fit. In general the work here clearly demonstrate the usefulness of the multinomial approximation in a real world setting.

4.7.2 The result

Given the data provided by Hare et al. (1998) and appropriate constraints, our results suggest that model 2 has a significantly better fit than model 1. This suggests that death in the cultures does not occur at the DP stage. This result was consistent for all data sets provided by Figures 5 and 6 (Hare et al., 1998). The results of our examination of the data provided by Figure 3 (Hare et al., 1998) were also consistent with these results, provided we examined the data from day 2 and day 3 independently. The results of a population test that assessed whether it was acceptable to combine data from more than one day, suggested that we were unable to do so.

Our best fitting model (Model 2) does not fit the data exactly. In biological situations exact fitting models are rare and this is mainly due to the inherent variation in biological systems. One could argue, therefore, that our results may also show that neither model sufficiently describes what actually takes place in the cultures. In reviewing the fits produced by Model 2 we see that the poorest correlations to the data are those to categories with the highest division number. This might suggest that the rates of division or death may be increasing or decreasing between generations respectively. Perhaps, those that survive the early stages of death and division are no longer being negatively selected and therefore are able to divide without being selected. Alternatively, those cells that survive negative selection may be dividing much more quickly.

In addition, higher division categories may be subject to greater inaccuracy with regard to cell counts. This is due to the poorer definition of higher division fluorescence peaks in the original data. However, it may be that the value of the likelihood function is relatively insensitive to the data in the higher division categories. A mathematical explanation might be that the value of the likelihood function itself may be dominated by the lowest division terms. Generally the lowest numbers of cells in any division category are those found with the highest division numbers. If we take into consideration the normalization procedure whereby these higher division categories are subject to division by the highest powers of 2, it would seem plausible that the overall value of the function is more dependent on the low division number categories than the higher. This may result in a skewing of the fit towards the low division categories.

However, such hypotheses as those described above can only be tested with the inclusion of additional parameters. More complex models would probably require more data. They may even be testable with the sort of multiple data sets provided Figure 3 Hare et al. (1998). However, it is doubtful whether more complex models would work with the data provided by Figures 5 and 6. This would leave us in the position of selecting hypotheses based on the nature of the data provided and would confuse matters. We therefore chose to limit our analysis to our original hypotheses.

Further to the above, we point out that given the nature of the data and the basic assumptions made about DP and SP cell behaviour further simple hypotheses are limited. Indeed, our results were not changed by increasing the number of parameters through application of the unconstrained general model (UGM) (section 4.1: figure 4.1). In all cases, parameter values obtained under the UGM were identical to those obtained using model 2, the best fitting constrained model (see section 4.4.3).

4.7.3 SP cell maturation

It has been observed SP cells are subject to a maturation that involves the regulation of two surface markers: heat stable antigen (HSA) and Qa-2. Immature SP cells are similar to DP cells in that they are $HSA^+/Qa\text{-}2^-$ and these cells have been shown to be susceptible to negative selection (Kishimoto and Sprent, 1997). Also, it is maintained that mature $HSA^-/Qa\text{-}2^-$ SP cells are subject to division (Sprent and Kishimoto, 2002). However, the data provided by Hare et al. (1998) suggests that division in FTOC cultures occurs directly after transfer to the SP stage. If we assume that mammalian cells take between 10 to 12 hours to divide and that after 3 days of culture some cells have divided at least 5 times it would seem plausible that some cells had entered into cycle immediately after transition to the SP stage. In conjunction with the observation that SP maturation takes several days this suggests that in these cultures immature SP cells enter into cycle.

This said, the introduction of a non-dividing SP stage, where SP cells could either transit to the dividing SP stage, die or do nothing, qualitatively did not change our results. The best fit provided by this model occurred when death in the first 2 stages was zero (result not shown). This is probably because this more complex model is subsumed by the UGM. Since two adjacent non-dividing stages can be combined into one overall stage. Indeed, interpretation of the more complex model is difficult because even if death did appear to occur at the second non-dividing stage it would be hard to state with certainty that this was actually an SP stage. It could just as well be a secondary DP stage during which cells are susceptible to death.

4.7.4 Modelling Death

The modelling contained herein does not explicitly model death. Cell death is inferred from the proportion of live cells. More precisely, we actually infer the proportion of CFSE dye that is lost to the experimenter. The loss of dye in these experiments can be attributed to either cell death or incomplete sampling. The latter would occur if the number of live cells obtained from the culture fell short of their actual number. The robustness study showed that, regardless of the proportion of live cells harvested from the culture, our results stand (figure 4.18).

4.7.5 Death and Negative Selection

Given our results, can any conclusions be drawn regarding the timing of negative selection? In examining this issue it is important that we discriminate between two types of cell death. Firstly, cells in these re-aggregate cultures may die for non-specific reasons, such as cell stress due to experimental manipulation. Alternatively, cells may undergo apoptosis because they have been negatively selected.

More specifically, re-aggregate cultures cannot be considered to emulate thymic conditions *in vivo* since thymic architecture is not replicated. However, Penit and Vasseur (1997) found evidence that SP expansion occurs *in vivo* through experiments using injections of BrdU or the anti-mitotic agent demecolcin. The question of whether results from *in vitro* protocols can be generalized to the biological situation *in vivo* is particularly relevant in the case of those data provided by Hare et al. (1998). This is because their experimental protocol calls for the depletion of hemopoietic cells during stromal preparation. This means that dendritic cells (DC), one of the most potent presenters of antigen in the thymus, would either be absent or in short supply. This depletion would therefore impact on the degree negative selection that is able to take place.

Indeed, Anderson et al. (1998) suggest that DC are by far the most effective APCs requiring only

the addition of 1% DC in order to achieve maximal 80% deletion. However, thymic stroma contains thymic epithelial cells and these are regarded as important APCs (Palmer, 2003). Interaction with these cells would therefore deplete auto-reactive cells. In addition thymocyte-thymocyte interaction may also play a role in negative selection (Choi et al., 2005). Murine thymocytes only express Class I MHC, however, so the selection of $CD4^+$ cells is not possible through this means. However, taken together these remarks would suggest that at least some of the death seen in these cultures is due to negative selection.

It may be possible to adapt the method used to obtain the data in Figure 6 (Hare et al., 1998) to obtain some additional information that may prove useful. The experimental protocol calls for matched and mismatched stroma. One would expect that negative selection to be lower in the latter of these preparations. Since the we would expect the reactivity of thymocytes positively selected on H-2^d stroma to be higher for matched than for mismatched stroma. However, this result is confounded by the fact that we do not know the sample size (the proportion of cells harvested from those available in the culture). If the sample size is identical then results suggest that, regardless of the model used, the probabilities of death are similar (figure 4.20), with mismatched stroma resulting in the slightly higher probability of death. However, this result can be reversed if the proportions are not identical. For example using model 2 we see that, assuming our H-2^b sample to be 100% of that available, the relative level of the probability of death can be reversed if the sample of H-2^d thymocytes was around 25% less.

In reference to this last point, a measure of sample size may be achievable with the addition to the culture of a fixed number of glass or polythene beads. Providing the cultures are well mixed assaying the proportion of glass beads found in the harvested sample would provide a measure of sample size. Several factors would have to be considered here. For example the size of glass bead would have implications with regard to accuracy. Adherence of cells to overly large beads may interfere with results. Indeed, a difficulty often associated with harvesting cells from culture is cell/cell adherence. Thymocytes interacting strongly with APC would be difficult to separate and this may also affect results (personal communication M. Dallman).

4.7.6 DP cells and death

Our results suggest that death does not occur at the DP $CD69^+$stage. This is puzzling and would appear at first sight to indicate that DP $CD69^+$ cells are resistant to death whatever the cause. This contradicts the observations of those that have observed negative selection in DP cells. However, these studies are generally based on transgenic cell lines. As mentioned previously, these are known to express high levels of TCR prior to the time when they would naturally occur. This would increase the probability of negative selection prior to a time when it would be likely to occur in

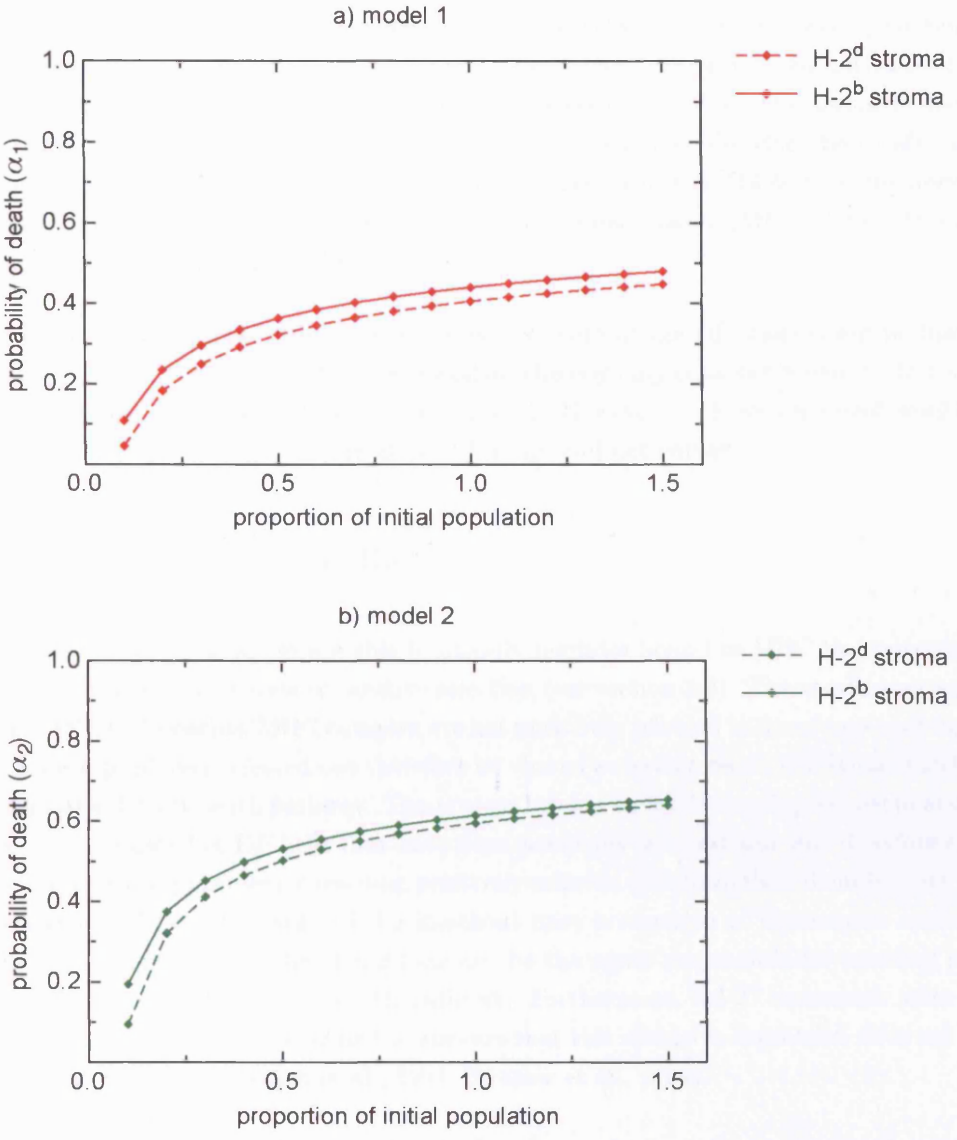

Figure 4.20: The effect of varying the initial population on the probabilities of death. The figure shows how the probabilities of death derived from data provided by Figure 6: H-2^d and H-2^b thymic stroma (Hare et al., 1998) using a) Model 1 ($\alpha_1 = 1 - \beta - \delta_1$) and b) Model 2 ($\alpha_2 = 1 - \gamma - \delta_2$) alter when the initial population is varied. The initial experimental population was 8×10^5 thymocytes.

wild type mice. This lends weight to the arguments of the proponents of the medulla as the site of negative selection as the observed DP deletion can be regarded as artifactual.

Baldwin et al. (1999) show that thymocytes may be liable to negative selection regardless of whether positive selection has already taken place. If negative selection were to occur just prior to or simultaneously with positive selection, CD69 expression would not only be an indicator of positive selection but also be might be indicative of negative selection survival. The argument here would be that Hare et al. (1998) used only CD69+ DP cells and may therefore be conducting their experiments on post-negatively selected cells. However, Baldwin et al. (1999) used thymocytes with TCR that are specific and have high affinity for moth cytochrome c (MCC)/I-Ek. The observed effect may therefore be non-physiological.

Yet another alternative explanation of the absence of death at the DP stage might be that experimental manipulation may incur a form of selection whereby only cells hardy enough to survive the initial preparatory steps are destined to be cultured. However, such an argument would require some explanation of why death occurs at the SP stage and not earlier.

4.7.7 The effect of bcl-2 and IL-7

In order ensure that thymocytes are able to identify peptides bound to MHC the naive thymocyte repertoire undergoes the process of positive selection (see section 2.2). Those cells that are unable to interact with self-peptide/MHC complex are not positively selected and undergo apoptosis. Cells that have been positively selected can therefore be viewed as having received a signal which enables them to avoid a default death pathway. The protein bcl-2 is an inhibitor of apoptosis in mammalian cells. It is up-regulated in DP cells that have been positively selected and was therefore suspected as being the agent responsible for rescuing positively selected cells from their default death pathway (Jameson et al., 1995). However in bcl-2 knockout mice production of thymocytes appears to be normal. This would indicate that bcl-2 may not be the agent responsible for enabling positively selected cells to avoid the default death pathway. Furthermore, bcl-2$^+$ transgenic mice produce thymocytes that over-express blc-2 and it appears that this excessive expression does not interfere with negative selection (Sentman et al., 1991; Strasser et al., 1994).

This latter result is important for 2 reasons. Firstly, in the experiments where the subjects are bcl-2 transgenic mice, it allows the possibility that some death to be due to negative selection. Secondly, Hare et al. (1998) propose and subsequently provide supporting evidence (Hare et al., 2000) suggest that IL-7 may be the factor responsible for inducing SP cells to divide. This has a possible knock on effect in that, IL-7 signalling is known to increase cell viability and this is probably through up-regulation of bcl-2 (Akashi et al., 1998). Provided bcl-2 expression has no effect on a cell's capacity to be negatively selected we may therefore conclude that SP cells not

only have an increased propensity to divide and have greater viability but are also open to negative selection.

To some extent increased viability of SP cells due to transgenic bcl-2 expression would suggest that death in cultures that used these cells would be more likely to be due to negative selection than non-specific causes. The fact that IL-7R expression also increases bcl-2 expression also means that in non bcl-2 transgenic cultures a similar effect would be seen. Overall then this supports our view that some death is due to negative selection and is not entirely due to non-specific causes.

4.7.8 In Conclusion

In conclusion our results suggest that cells death in the FTOC cultures conducted by Hare et al. (1998) does not occur at the DP 69^+ stage. Death therefore appears to occur at the SP stage and may happen in conjunction with division. In reaching this conclusion we have inferred the behaviour of the cells from that of the CFSE dye. In the case of cell death this inference is obtained through dye loss. However, dye loss can occur in 3 ways: cells can either die from non-specific causes, be negatively selected or be the result of incomplete harvesting of thymocytes. In the latter case our robustness study indicates that incomplete harvesting would not affect our result.

However we cannot tell whether death is due to negative selection or non-specific causes since we are unable to separate these using our models. However, given that the cultures contain some types of APC' we would expect that part of this death was due to negative selection. Indeed, the proportion of death due to negative selection may be high as some of the experiments were conducted using bcl-2 transgenic cells. These cells are resistant to death through non-specific causes but have been shown to be susceptible to negative selection. In addition cells expressing IL-7R are not only stimulated by IL-7 to divide but also express bcl-2. This suggests that non-specific death in non-bcl-2 transgenic cells would also be low and subsequently points to their death being caused by negative selection. These subsidiary observations therefore add support to the idea that negative selection is taking place in these cultures, at the SP stage and may be accompanied by division. This also supports the view that negative selection would occur in the medulla since this is where SP cells reside prior to export to the periphery.

The work here clearly demonstrates the usefulness of our multinomial approximation to the likelihood in a real world context. We have also demonstrated that when using the multinomial approximation our inability to use standard methods for obtaining CIs, and conducting the likelihood ratio test, can be overcome through Monte-Carlo methods.

Chapter 5

The Continuous Time Model

5.1 Introduction

Thus far we have used a discrete time branching process model to examine the data published by Hare et al. (1998). In this type of model all cells are assumed to undergo some kind of decision making process at each time step. For example, our SP cells decided to either die, divide or do nothing at each time step. We can view this assumption as a form of synchronization in which each decision event is made in a unit of time. This can be considered to model the relative impact of being born to a particular generation; later generations have less opportunity to make decisions.

However, real cells do not behave identically with respect to division cycle length. This behavioural heterogeneity is loosely approximated in a 3 way branching process. For example, a cell that divides twice in 2 time steps could be considered to be dividing twice as quickly as a cell that has only divided once in the same number of steps. However, the degree to which heterogeneity of cycle length can be represented in a discrete time model is dependent upon and limited by the number of time steps used. Reusing our example above, if our model contained a maximum of 2 time steps our dividing population would only be represented by two types of cells, one dividing twice as fast as the other.

Increasing the number of division steps in the model would therefore appear to be an answer to this problem. However, the mathematical method for deriving the likelihood function in the simplest discrete case involves raising a $k+1$ by $k+1$ matrix to the kth power (where k is the number of types). This approach would therefore become computationally cumbersome for large k. In addition, our parameters have values that lie between zero and one. If k were large the possibility of rounding errors caused by computing the raising of a parameter to the kth power would also

have to be considered.

Given these difficulties we asked the question: would it be possible to use a continuous time model? In addition, we naturally asked would the increased resolution afforded by such an approach result in a better fitting model? Here we therefore show how to derive the continuous time model. We subsequently repeat some of our analysis of the data provided by Hare et al. (1998) using a constrained general model. This once again tests hypotheses 1 and 2 (see section 4.1.4). The results suggest that regardless of which constraints are imposed the continuous time model is a poorer model of the data than the discrete model.

5.2 Mathematical Methods

5.2.1 The Time-Continuous Branching Process

The principal reference for the following paragraphs is Kimmel and Axelrod (2002): chapters 4 and 5.

The general form of the time continuous branching process is the Bellman-Harris process. In this process a "particle" is born at time zero and lives for a time τ; a random variable with cumulative distribution function $G(\tau)$. At the end of its "lifetime" the particle ceases to exist and gives rise to a number of progeny $n \in \{0, 1, 2 \ldots\}$. The probability that there are n progeny is provided by the progeny probability generating function $f(s)$. The behaviour of the progeny is independent of that of the ancestor and each other, but is similar to that of the ancestor in that their lifetimes are distributed with $G(\tau)$ and the probabilities of each producing n offspring dictated by the progeny pgf $f(s)$. The entire process has the probability generating function $F(s,t)$ and continues provided that at time $t \geq 0$ the number of particles $Z(t) \neq 0$.

This "age-dependent" process is clearly stochastic as it depends upon $G(\tau)$ and $f(s)$. However, Bellman-Harris processes are often hard to analyse and are generally non-Markovian except for two special cases: the Galton-Watson process and the Markov age dependent process. In the case of the Galton-Watson process all lifetimes are identical, thus making the process discrete. In the Markov age dependent process lifetimes are exponentially distributed ie. in terms of the cdf $G(\tau) = 1 - e^{-\lambda\tau}$.

Here we use the latter case, since the use of exponentially distributed lifetimes often leads to results amenable to analysis. In addition, the results achieved under the exponential assumption can be used as a basis on which to "conjecture" the properties of more complex models. Within the context of modelling cells, the exponential assumption has a weakness in that it implies that there lifetimes that the cell cycle time can be arbitrarily short. In a given time, this allows finite possibilities of

physiologically impossible numbers of divisions to exist. However, the results of maximum likelihood estimation are driven by the data and provided the model is in some sense correct the probabilities of unphysiological numbers of divisions should be small enough to be ignored.

5.2.2 The Differential Equation of the Process pgf

Consider a cell that starts life at time 0. This cell has two behavioural traits. Firstly, it has a lifetime τ which is exponentially distributed with parameter λ. Secondly, at the end of its life the cell produces a number of progeny $n \in \{0, 1, 2\}$ with a probability distribution with pgf $f(s)$. Each first generation progeny acts independently but inherits the ancestor cell's traits and this behaviour is repeated for all subsequent generations. Due to the stochastic nature of this process the number of cells that are alive at time $t \geq 0$ is therefore a random variable Z_t. The probability that at time t the number of cells $Z_t = z$ where $z \in \{0, 1, 2 \ldots\}$ is obtained from the process pgf $F(s,t)$ (Kimmel and Axelrod, 2002).

The process pgf satisfies a partial differential equation and we follow Kimmel and Axelrod (2002) in its derivation [1]. First we note that at any point in time $t \geq 0$ any living cell will have its remaining lifetime distributed exponentially with parameter λ. This being a consequence of the no memory property of the exponential distribution. Each of these cells can therefore be thought of as the ancestor cell of a subprocess, started at time t, that is independently and identically distributed (iid) to the entire process started by the original ancestor at time 0. Therefore, at a time $t + \Delta t$ the number of cells present $Z_{t+\Delta t}$ will be equal to the sum of all cells present in the iid subprocesses started at the earlier time Δt. We therefore have

$$Z_{t+\Delta t} = \sum_{i=1}^{Z_{\Delta t}} Z_t^i \tag{5.1}$$

where the ith iid subprocess is indicated by the superscript i. In terms of the process pgf this can be redefined as

$$F(s, t + \Delta t) = F[F(s,t), \Delta t] \tag{5.2}$$

[1] We deviate from Kimmel and Axelrod (2002) by adopting the use of partial derivative which enables us to keep $F(s,t)$ as a function of two variables. Kimmel and Axelrod (2002) alter the form of $F(s,t)$ to be a function of s with parameter t ie. $F(s;t)$. However the aim here is to derive the differential of $F(s,t)$ with respect to t. The notation of Kimmel and Axelrod (2002) is therefore slightly confusing even if their meaning is clear.

The initial condition for our process is that it is started by one cell. Therefore at time 0 we have $F(s,0) = s$. Using this and subtracting $F(s,t)$ from both sides we obtain

$$F(s,t+\Delta t) - F(s,t) = F[F(s,t),\Delta t] - F[F(s,t),0] \tag{5.3}$$

We now note that if Δt is small the process will include, with probability close to 1, either the ancestor or its first generation progeny.

$$F(s,\Delta t) = s e^{-\lambda \Delta t} + f(s)(1 - e^{-\lambda \Delta t}) + o(\Delta t) \tag{5.4}$$

or alternatively

$$F(s,\Delta t) - F(s,0) = \{-s + f(s)\}(1 - e^{-\lambda \Delta t}) + o(\Delta t) \tag{5.5}$$

substituting from equation 5.3 therefore have

$$\frac{F(s,t+\Delta t) - F(s,t)}{\Delta t} = \frac{\{-F(s,t) + f[F(s,t)]\}(1 - e^{-\lambda \Delta t}) + o(\Delta t)}{\Delta t} \tag{5.6}$$

Letting $\Delta t \to 0$ therefore produces the partial differential equation

$$\frac{\partial F(s,t)}{\partial t} = -\lambda \{F(s,t) - f[F(s,t)]\} \tag{5.7}$$

This can be extended for a general process where there are multiple types of particle k, numbering $0, 1, \ldots, \eta$ (Kimmel and Axelrod, 2002). Denoting $\mathbf{s}$ as the vector of dummy variables s_k, each type would produce offspring according to $f_k(\mathbf{s}) = f_k(s_0, s_1, \ldots, s_\eta)$. In addition the lifetime of each individual of type k would be exponentially distributed with parameter λ_k. Further denoting $\mathbf{F}$, $\mathbf{f}$ and λ_v as vectors containing the process and progeny pgfs in addition to our λs for each type

respectively, we find

$$\frac{\partial \mathbf{F}(\mathbf{s},t)}{\partial t} = -\lambda_v \cdot \{\mathbf{F}(\mathbf{s},t) - \mathbf{f}[\mathbf{F}(\mathbf{s},t)]\} \tag{5.8}$$

with initial conditions $F_k(s_k,0) = s_k$. Here the dot operator denotes a scalar product of 2 vectors.

5.2.3 Time Continuous Modelling of CFSE Distribution

We proceed by assuming that at the end of a lifetime a unit of dye either transitions to the $k+1th$ division state or is lost with probabilities γ and $\alpha = 1-\gamma$ respectively. If we define a cell type to be synonymous with the division state it inhabits we can form a continuous time multitype model as described above. In terms of the progeny pgf we therefore have $f_k(\mathbf{s}) = \alpha + \gamma s_{k+1}$. In addition, for simplicity, we will also assume all lifetimes to be identically distributed ie. $\lambda_k = \lambda$. We also assume that all cells of the type η only give birth to cells of type η. Denoting $F_k = F_k(s_k,t)$ from above we therefore obtain the cascade system of PDEs:

$$\frac{\partial F_k}{\partial t} = -\lambda F_k + \lambda[\alpha + \gamma F_{k+1}] \tag{5.9}$$

for $k = \eta$ we have

$$\frac{\partial F_\eta}{\partial t} = -\lambda F_\eta + \lambda[\alpha + \gamma F_\eta] \tag{5.10}$$

Using the integrating factor $e^{\lambda t}$ this system of equations can easily be solved by back substitution yielding

$$F_\eta = 1 + e^{-\lambda\alpha t}(s_\eta - 1) \tag{5.11}$$

and for $k < \eta$

$$F_k = 1 + e^{-\lambda\alpha t}(s_\eta - 1) + e^{-\lambda t}\sum_{i=0}^{\eta-k}(s_{i+k} - s_\eta)\frac{(t\lambda\gamma)^i}{i!} \tag{5.12}$$

If we start with one unit of dye at in state zero, the expectation $E[Z_k]$ of dye being found in state k is obtained by differentiating F_0 with respect to s_k and subsequently setting all $s_k = 1$. Thus we obtain

$$E[Z_\eta] = e^{-\lambda\alpha t} - e^{-\lambda t}\sum_{i=0}^{\eta-1}\frac{(t\lambda\gamma)^i}{i!} \tag{5.13}$$

and for $k < \eta$

$$E[Z_k] = \frac{e^{-\lambda t}(t\lambda\gamma)^k}{k!} \tag{5.14}$$

The biological explanation of these expectations is that equation 5.14 gives us the expectation of dye belonging to any given division state k. Equation 5.13 gives us the expectation of the total amount of dye found in all division states greater than or equal to η. These expectations can be used in a various ways to model the data depending on how we wish to model it (see below). However, the process of forming an approximate log-likelihood from these expectations is essentially the same as that used for the discrete model.

5.2.4 Three Ways of Modelling the Data

One difficulty in interpreting CFSE data is uncertainty about the maximum number of divisions that have occurred during the experiment. This uncertainty is mainly due to the level of background noise that is often present in CFSE profiles. The experimentalist can only extract the peaks that are observable above this noise. This may result in truncation of the data and this under-represents what has occurred. A further possibility is that more divisions may have occurred during incubation but the cells that had undergone greater than the observed number of divisions died.

These possibilities have an impact on how we attempt to model the data. They present us with 3 options:

1. Assume that the data is correct and only the number of divisions present are those that took place.
2. Assume more divisions took place but the extra cells created died.
3. Assume more divisions took place but the data was truncated due to difficulty in identifying peaks.

In modelling option 1 we require only to know the expectations of the dye in each division category found in the data. This can be obtained from equation 5.14 above. From this, we can obtain the expected quantity of dye that has disappeared $E[Z_l]$:

$$E[Z_l] = 1 - \sum_{k=0}^{\eta-1} E[Z_k] \tag{5.15}$$

where η is one greater than the observed number of divisions.

In modelling option 2 we can assume that the expectations of the dye found in each division state greater than those found in the data to be given by equation 5.14. We would then proceed by increasing the number of possible divisions to some number $\nu \geq \eta$ and include the expectations of these in our calculations. The expectation of the total amount of dye found in divisions states greater than those observed as $\nu \to \infty$ is given by equation 5.13. In the first instance the expectation of dye lost is therefore

$$E[Z_{l,\nu}] = 1 - \sum_{k=0}^{\nu} E[Z_k] \tag{5.16}$$

where on the LHS the subscript l, ν indicates our dependence on ν and in the limit $\nu \to \infty$ this becomes

$$\lim_{\nu \to \infty} E[Z_{l,\nu}] = E[Z_{l,\infty}] = 1 - E[Z_\eta] - \sum_{k=0}^{\eta-1} E[Z_k] \tag{5.17}$$

Modelling option 3 requires that we employ some method of estimating our missing data. Here

we use the Expectation Maximization, or EM, algorithm (Little and Rubin, 1987). Briefly, this proceeds by first estimating on the observed data. Next, we calculate the expectations of the missing data from the obtained MLEs. We now re-estimate on the observed data augmented with our expectations of the missing data. We subsequently use the MLEs obtained from this re-estimation to calculate a revised set of expectations for the missing data. The process is re-iterated until the estimates converge. Here we can use $E[Z_k]$ (equation 5.14) to calculate the expectations of missing data belonging to division states $k \in \{\eta, \eta+1, \eta+2, \ldots, \nu\}$ with the expectation of lost dye calculated using $E[Z_{l,\nu}]$ (equation 5.16). In the limit $\nu \rightarrow \infty$ we can use $E[Z_{l,\infty}]$ (equation 5.17) to determine the expectation of the total quantity of dye lost. In this case, the dye found in division states greater than or equal to η we therefore use $E[Z_\eta]$ (equation 5.13) and for division states less than η we use $E[Z_k]$ (equation 5.14).

5.2.5 Derivation of the Model for CFSE data

In order to examine the timing of death in the FTOC cultures of Hare et al. (1998) we construct a general model with assumptions:

1. DP cells have lifetimes that are exponentially distributed with parameter λ_1. Upon reaching the end of their lifetime each cell either transitions to the SP stage with probability β or dies with probability $\alpha_1 = 1 - \beta$.
2. SP cells also have exponentially distributed lifetimes with parameter λ_2. At the end of their lives these cells either divide with probability γ or die with probability $\alpha_2 = 1 - \gamma$.

We yet again take the approach of modelling the distribution of dye rather than the distribution of cells. We denote the process pgf for DP cells as $F = F(s,t)$ and for SP cells $F_k = F_k(s,t)$ where k is the division state of single positive cells. Conditioning on the presence of dye, for DP cells we therefore have

$$\frac{dF}{dt} = -\lambda_1 F + \lambda_1[\alpha_1 + \beta F_0] \tag{5.18}$$

Whilst for SP cells we have

$$\frac{dF_k}{dt} = -\lambda_2 F_k + \lambda_2[\alpha_2 + \gamma F_{k+1}] \tag{5.19}$$

for $k = 0, \ldots, \eta - 1$.

Thus equations 5.18, 5.19 form a cascade system of differential equations with initial conditions to $F(o) = s$ and $F_k(0) = s_k$. As above, working progressively in reverse order we solve this system for F_k and F using the respective integrating factors: $e^{\lambda_2 t}$ and $e^{\lambda_1 t}$. We therefore have the solutions:

$$F_k = 1 + e^{-\lambda_2 \alpha_2 t}(s_\eta - 1) + e^{-\lambda_2 t} \sum_{i=0}^{\eta-k} (s_{i+k} - s_\eta) \frac{(t\lambda_2 \gamma)^i}{i!} \tag{5.20}$$

and

$$F e^{\lambda_1 t} = \beta \lambda_1 \int e^{\lambda_1 t} F_0 dt + C \tag{5.21}$$

taking the integral term as a function $g(t)$, we have

$$g(t) = \int e^{\lambda_1 t} + e^{t(\lambda_1 - \lambda_2 \alpha_2)}(s_\eta - 1) + e^{t\epsilon} \sum_{i=0}^{\eta} (s_i - s_\eta) \frac{(t\lambda_2 \gamma)^i}{i!} dt \tag{5.22}$$

where $\epsilon = \lambda_1 - \lambda_2$. Integrating we obtain,

$$g(t) = \frac{e^{\lambda_1 t}}{\lambda_1} + \frac{e^{t(\lambda_1 - \lambda_2 \alpha_2)}(s_\eta - 1)}{\lambda_1 - \lambda_2 \alpha_2} + e^{t\epsilon} \sum_{k=0}^{\eta} \sum_{i=0}^{k} (-1)^{i-k} (s_k - s_\eta) \frac{(\lambda_2 \gamma)^k t^i \epsilon^i}{i! \epsilon^{k+1}} \tag{5.23}$$

substituting this in 5.21 and remembering that at $F(0) = s$ we find

$$C = s - \beta \lambda_1 g(0) = s - \beta \lambda_1 \left\{ \frac{1}{\lambda_1} + \frac{(s_\eta - 1)}{\lambda_1 - \lambda_2 \alpha_2} + \sum_{k=0}^{\eta} (-1)^k (s_k - s_\eta) \frac{(\lambda_2 \gamma)^k}{\epsilon^{k+1}} \right\} \tag{5.24}$$

substituting into 5.21 we now have have

$$F = s\epsilon^{-\lambda_1 t} - \epsilon^{-\lambda_1 t}\beta\lambda_1$$

$$\left\{\frac{1-\epsilon^{\lambda_1 t}}{\lambda_1} + \frac{(s_\eta - 1)(1-\epsilon^{t(\lambda_1 - \lambda_2\alpha_2)})}{\lambda_1 - \lambda_2\alpha_2} + \sum_{k=0}^{\eta}(s_k - s_\eta)(\lambda_2\gamma)^k[\frac{(-1)^k}{\epsilon^{k+1}} - \epsilon^{t\epsilon}\sum_{i=0}^{k}(-1)^{i-k}\frac{t^i\epsilon^i}{i!\epsilon^{k+1}}]\right\}$$

Denoting the expectation of dye being found in a DP cell as $E[Z]$ and in SP division state k as $E[Z_k]$ we differentiate F with respect to s or s_k and evaluate at $s_k = 1$ to obtain

$$E[Z] = \epsilon^{-\lambda_1 t} \tag{5.25}$$

$$E[Z_k] = -\epsilon^{-\lambda_1 t}\beta\lambda_1(\lambda_2\gamma)^k\left\{(-1)^k\frac{1}{\epsilon^{k+1}} - \epsilon^{t\epsilon}\sum_{i=0}^{k}(-1)^{i-k}\frac{t^i\epsilon^i}{i!\epsilon^{k+1}}\right\} \tag{5.26}$$

for $k < \eta$. Whilst for $k = \eta$ we have

$$E[Z_\eta] = -\epsilon^{-\lambda_1 t}\beta\lambda_1\left\{\frac{1-\epsilon^{t(\lambda_1-\lambda_2\alpha_2)}}{\lambda_1 - \lambda_2\alpha_2} + \sum_{k=0}^{\eta-1} -(\lambda_2\gamma)^k[\frac{(-1)^k}{\epsilon^{k+1}} - \epsilon^{t\epsilon}\sum_{i=0}^{k}(-1)^{i-k}\frac{t^i\epsilon^i}{i!\epsilon^{k+1}}]\right\} \tag{5.27}$$

From equation 5.26 this simplifies to

$$E[Z_\eta] = \epsilon^{-\lambda_1 t}\beta\lambda_1\left\{\frac{\epsilon^{t(\lambda_1-\lambda_2\alpha_2)}-1}{\lambda_1-\lambda_2\alpha_2}\right\} - \sum_{k=0}^{\eta-1}E[Z_k] \tag{5.28}$$

Returning to equation 5.26 and taking ϵ^{k+1} as a common denominator, for $E[Z_k]$ we obtain

$$E[Z_k] = -\frac{e^{-\lambda_1 t}\beta\lambda_1(\lambda_2\gamma)^k\left\{(-1)^k - e^{t\epsilon}\sum_{i=0}^{k}(-1)^{i-k}\frac{t^i\epsilon^i}{i!}\right\}}{\epsilon^{k+1}} \tag{5.29}$$

the solution in the form of equation 5.29 is numerically unusable. This is because as $\epsilon \to 0$ the denominator also tends to 0. We therefore take the bracketed terms and using the series expansion of $e^{t\epsilon}$ we obtain

$$\begin{aligned}(-1)^k - e^{t\epsilon}\sum_{i=0}^{k}(-1)^{i-k}\frac{t^i\epsilon^i}{i!} &= (-1)^k - \sum_{i=0}^{\infty}\frac{t^i\epsilon^i}{i!}\sum_{i=0}^{k}(-1)^{i-k}\frac{t^i\epsilon^i}{i!} \\ &= \sum_{i=0}^{\infty} -\frac{t^{i+k+1}\epsilon^{i+k+1}}{k!i!(i+k+1)} \\ &= t^{k+1}\epsilon^{k+1}\sum_{i=0}^{\infty} -\frac{t^i\epsilon^i}{k!i!(i+k+1)}\end{aligned}$$

substitution into equation 5.29, further simplification yields

$$E[Z_k] = \frac{e^{-\lambda_1 t}\beta\lambda_1(\lambda_2\gamma)^k t^{k+1}\sum_{i=0}^{\infty}\frac{t^i\epsilon^i}{i!(i+k+1)}}{k!} \tag{5.30}$$

In this form we find that

$$\lim_{\epsilon\to 0}\sum_{i=0}^{\infty}\frac{t^i\epsilon^i}{i!(i+k+1)} = \frac{1}{k+1} \tag{5.31}$$

We therefore no longer have the problems created by the denominator in equation 5.29.

5.2.6 The expectation of zero divisions

Note that in formulating our system of equations we have kept the expectations of zero divisions for DP cells $E[Z]$ and SP cells $E[Z_0]$ separate. We obtain the combined expectation of zero divisions overall $E[Z_0']$ by simply summing the two. For the general model we therefore have

$$E[Z_0'] = E[Z] + E[Z_0] = e^{-\lambda_1 t}\left\{1 + \beta\lambda_1 t \sum_{i=0}^{\infty} \frac{t^i \epsilon^i}{i!(i+1)}\right\} \tag{5.32}$$

5.2.7 Constraining the model

We constrain the model in two ways. For hypothesis 1, death only occurs at the DP stage, we set $\gamma = 1$ (model 1). For hypothesis 2, death occurs at the SP stage, we set $\beta = 1$ (model 2). Assuming that we combine expectations for zero divisions, we obtain the expectations: equation 5.32 with

$$E[Z_k] = \frac{e^{-\lambda_1 t}\beta\lambda_1\lambda_2^k t^{k+1}\sum_{i=0}^{\infty}\frac{t^i\epsilon^i}{i!(i+k+1)}}{k!} \tag{5.33}$$

and

$$E[Z_\eta] = \beta(1 - e^{-\lambda_1 t}) - \sum_{k=0}^{\eta-1} E[Z_k] \tag{5.34}$$

for model 1. For model 2 our constraint yields

$$E[Z_0'] = e^{-\lambda_1 t}\left\{1 + \lambda_1 t \sum_{i=0}^{\infty} \frac{t^i \epsilon^i}{i!(i+1)}\right\} \tag{5.35}$$

with

$$E[Z_k] = \frac{e^{-\lambda_1 t}\lambda_1\gamma^k\lambda_2^k t^{k+1}\sum_{i=0}^{\infty}\frac{t^i\epsilon^i}{i!(i+k+1)}}{k!} \tag{5.36}$$

and

$$E[Z_\eta] = e^{-\lambda_1 t}\lambda_1 \left\{ \frac{e^{t(\lambda_1 - \lambda_2\alpha_2)} - 1}{\lambda_1 - \lambda_2\alpha_2} \right\} - \sum_{k=0}^{\eta-1} E[Z_k] \tag{5.37}$$

The formation of the approximate log-likelihood function from the expectations follows the method laid out for the discrete models. We follow section 5.2.4 in order to obtain variations depending on how we view the data. Noting that equation 5.15 becomes

$$E[Z_l] = 1 - E[Z'_0] - \sum_{k=1}^{\eta-1} E[Z_k] \tag{5.38}$$

Equation 5.16 changes to

$$E[Z_{l,\nu}] = 1 - E[Z'_0] - \sum_{k=1}^{\nu} E[Z_k] \tag{5.39}$$

and equation 5.17 becomes

$$\lim_{\nu\to\infty} E[Z_{l,\nu}] = E[Z_{l,\infty}] = 1 - E[Z_\eta] - E[Z'_0] - \sum_{k=1}^{\eta-1} E[Z_k] \tag{5.40}$$

5.2.8 Minimization of the likelihood function

Minimization of the negative log-likelihood function was achieved using the simplex method of Nelder-Mead (Press et al., 2002). The stopping criteria for this algorithm was that the standard deviation of the function values calculated at the simplex vertices was less than 10^{-10}. Note that when attempting to use the method of global minimization proposed by Stanton et al. (1997) difficulties were encountered. This algorithm seemed to rapidly close to a location on the surface and then proceed very slowly towards the minimum. This suggests that the surface has a narrow "valley" through which its points are forced to squeeze in the direction of the minima. With tuning, the method did prove capable of finding the correct minima however the time taken was, for our purposes, excessive (hours rather than minutes).

5.3 Results

5.3.1 Adjusted Log-likelihoods

In Chapter 4 the approximate log-likelihoods we quoted were obtained without the use of the multinomial coefficient. This is acceptable because the likelihoods obtained using different models were based on identical data. In the following results our observed data does not change. However, in the case of the EM algorithm we estimate missing data. This means that the estimates of the missing data become model dependent. In order to offset this dependency we can incorporate the multinomial coefficient into our calculation of the likelihood. Subsequently, in this chapter we quote adjusted approximate log-likelihoods.

5.3.2 Estimates are unbiased

Initially, we test to see if using our continuous time models, the approximate likelihood method produces unbiased parameter estimates. For a chosen model, we therefore simulated data (1000 sets) using the parameter values estimated from the data using that model. A distribution of estimates was then derived from the simulated data sets. For each parameter, comparison of the simulated distribution mean with that of the original input value revealed that the estimates are unbiased (figure 5.1).

5.3.3 Results relating to Figure 6: H-2d thymic stroma (Hare et al., 1998): Option 1

For option 1 we assume that the observed data is all that exists and no other division events have occurred. The multinomial approximation has the form:

$$l(\boldsymbol{\theta} \mid data) = \log\left(\frac{N_s!}{d_0!d_1!\ldots d_{\eta-1}!d_l!}\right) + d_l \log E[Z_l] + d_0 \log E[Z_0'] + \sum_{i=1}^{\eta-1} d_i \log E[Z_i] \qquad (5.41)$$

where N_s is the initial experimental population of cells, d_i represents the dye found in division state i and d_l is the inferred quantity of dye lost obtained from the normalized observations and the initial population. In the case of the data provided by Hare et al. (1998) the maximum number of divisions observed is five we therefore set $\eta = 6$.

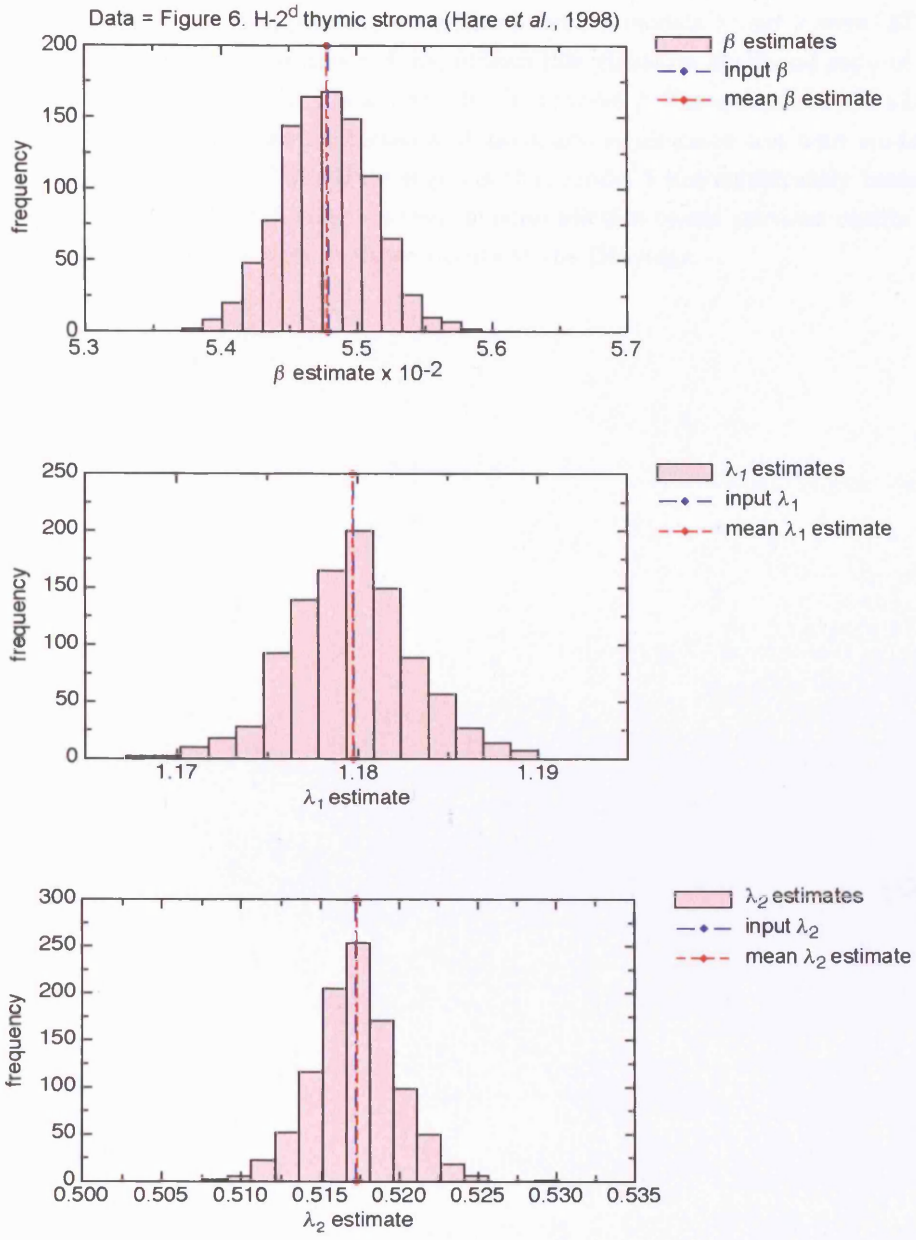

Figure 5.1: The continuous approximation produces unbiased parameter estimates. Continuous time model 1 (option 1.) parameter estimates (input MLEs:$\beta = .05478$, $\lambda_1 = 1.17972$, $\gamma = 1.0$ and $\lambda_2 = 0.51719$) for the data provided by Figure 6. H-2^d thymic stroma Hare et al. (1998) were used to create 1000 simulated data sets. The figure shows the distributions of Model 1 MLEs produced by estimating from these data sets. The mean of the distributions and is shown along with the input MLEs used for the simulation.

Following estimation using equation 5.41 comparison of the fits of model 1 and 2 to the data provided by Figure 6: H-2^d thymic stroma (Hare et al., 1998) suggests that either model may fit the data (figure 5.2). The approximate log-likelihood for models 1 and 2 were -87 and -129 respectively. With model 1 acting as a null hypothesis this yielded a likelihood ratio of -42. This negative value indicates that model 1 is a better fit than model 2. Since the choice of which model acts as a null is arbitrary here we conducted a Monte-Carlo significance test with model 2 acting as the null model (figure 5.3). This figure suggests that model 1 has significantly better fit than model 2 ($p < .001$). Overall result suggests that, in contradiction to our previous results under the discrete model, death in the culture analysed occurs at the DP stage.

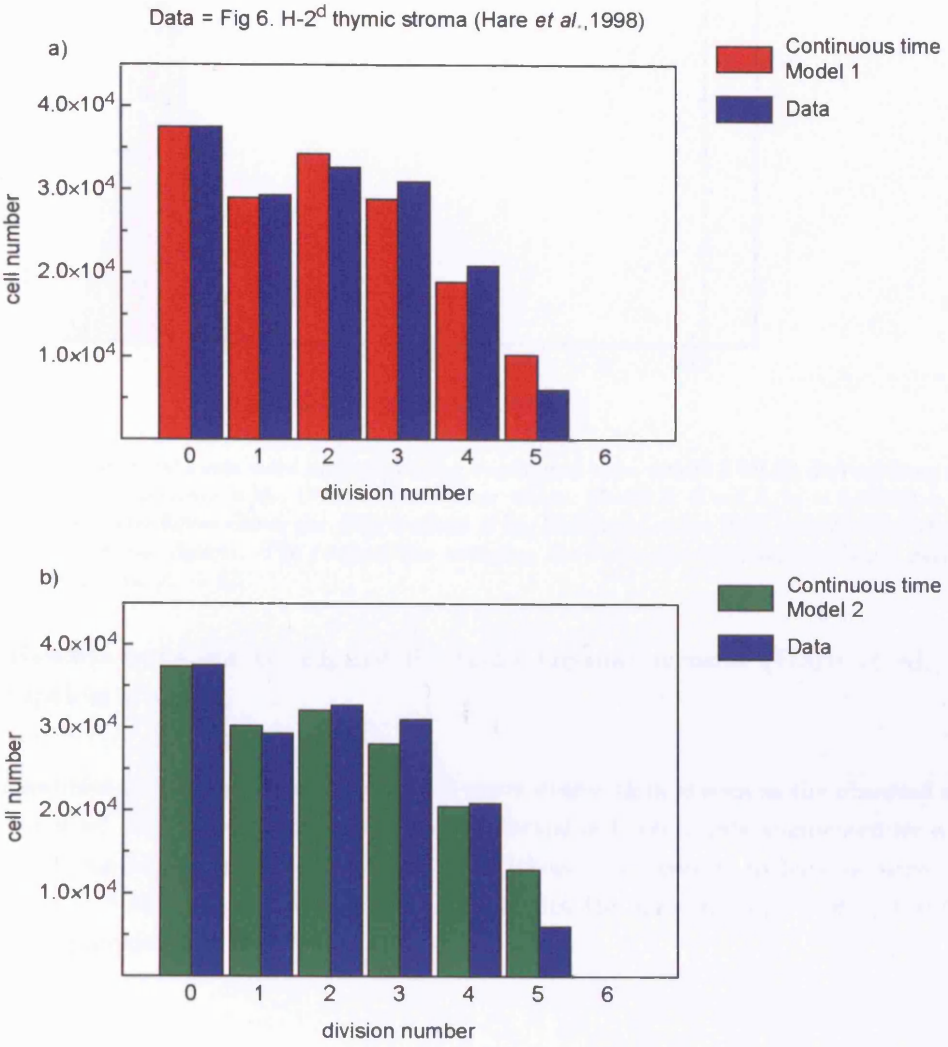

Figure 5.2: The mean cell count from 1000 simulations of continuous time a) model 1 (red) and b) model 2 (green) are shown with the experimental results from Figure 6. H-2^d thymic stroma (Hare et al., 1998) (blue). The simulated data is truncated and only includes up to 5 cell divisions. The parameter values used in the simulations were the MLEs derived from the experimental data; Model 1: $\beta = .05478$, $\lambda_1 = 1.17972$, $\gamma = 1.0$, $\lambda_2 = 0.51719$ and Model 2: $\beta = 1.0$, $\lambda_1 = 1.33832$, $\gamma = .32114$ and $\lambda_2 = 2.09153$.

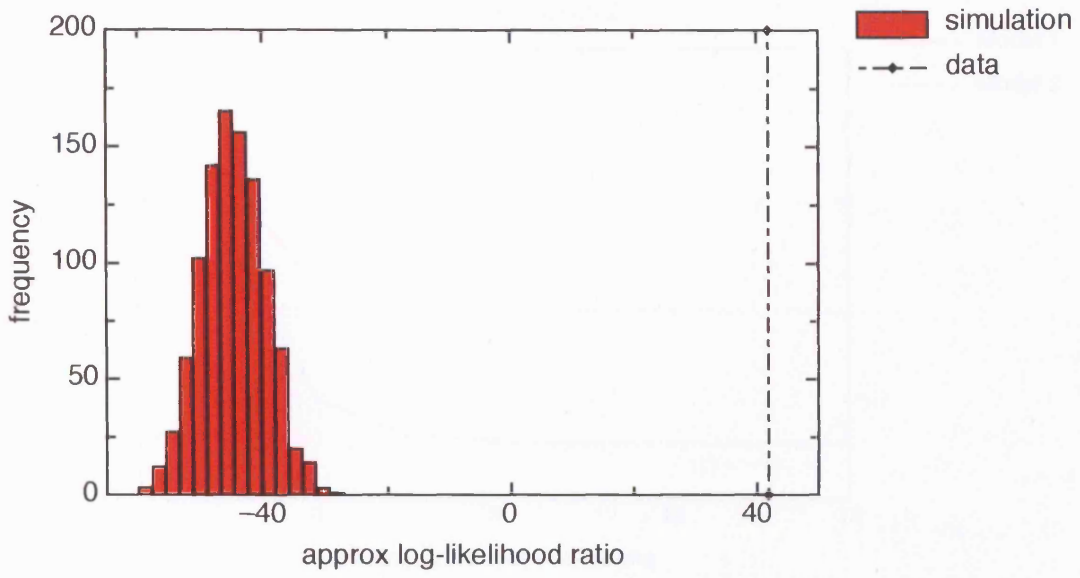

Figure 5.3: Simulated data sets were produced using continuous time model 2 MLEs derived from a) Figure 6: H-2^d thymic stroma (Hare et al., 1998) as parameter values: Model 2: $\beta = 1.0$, $\lambda_1 = 1.33832$, $\gamma = .32114$ and $\lambda_2 = 2.09153$. The figure shows the distributions of log-likelihood ratios (red) derived from these model 2 simulations (1000 per figure). The vertical line indicates the respective log-likelihood ratio derived from the experimental data ($\ell_r = 42$).

5.3.4 Results relating to Figure 6: H-2d thymic stroma (Hare et al., 1998): Option 2

Option 2 assumes that there have been more division events than is seen in the observed data but the cells involved died. In this case we fit to the normalized cell counts augmented by a number of zero counts up to the maximum number of divisions ν we assume to have occurred. In the case where $\nu = 8$ and $\eta = 6$ we would fit to data with the form $\{d_0, d_1, \ldots, d_{\eta-1}, 0, 0, 0\}$. The multinomial approximation has the form

$$l(\boldsymbol{\theta} \mid data) = \log\left(\frac{N_s!}{d_0!d_1!\ldots d_{\eta-1}!d_l!}\right) + d_l \log E[Z_{l,\nu}] + d_0 \log E[Z_0'] + \sum_{i=1}^{\eta-1} d_i \log E[Z_i] \quad (5.42)$$

and in the limit $\nu \to \infty$ ie. the number of zeros tends to infinity we have

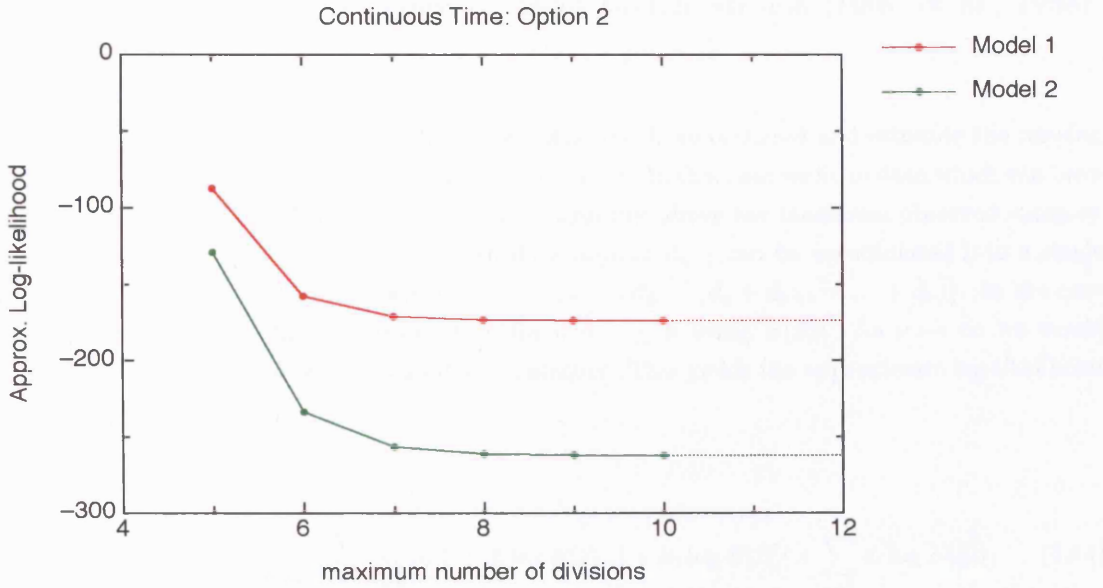

Figure 5.4: The effect of extending the data assuming that all cells that have undergone more than the observed number of divisions have died (option 2) on the continuous time models 1 and 2. The likelihoods are those obtained using the data provided by Figure 6 H-2^d thymic stroma.(Hare et al., 1998). The results using the continuous time general model (UGM) were identical to those obtained for model 1. For either model, the solid line is the result obtained by assuming the maximum number of divisions (ν) is finite and this rapidly approaches the assymptotic result: $\nu \to \infty$ (dotted lines).

$$l(\boldsymbol{\theta} \mid data) = \log\left(\frac{N_s!}{d_0!d_1!\ldots d_{\eta-1}!d_l!}\right) + d_l \log E[Z_{l,\infty}] + d_0 \log E[Z_0'] + \sum_{i=1}^{\eta-1} d_i \log E[Z_i] \quad (5.43)$$

The results of estimating on the data provided by Figure 6: H-2d thymic stroma (Hare et al., 1998) using both these equations are shown in figure 5.4. This figure shows that as the maximum number of divisions assumed to have occurred increases so the approximate log-likelihood monotonically decreases. Regardless of the model the decline in likelihood nears its assymptotic limit (estimation using equation 5.43) at around 8 divisions (dotted lines). These results suggest that option 2 would not be an appropriate way to model the data since the best fit is achieved on the observed data with no additional zero data. Overall the results therefore agree with the general continuous time result from option 1 that model 1 has a better fit than model 2.

5.3.5 Results relating to Figure 6: H-2d thymic stroma (Hare et al., 1998): Option 3

Here we assume that more divisions than those observed have occurred and estimate the missing data using the EM algorithm as described in section 5.2.4. In this case we fit to data which can have the form $\{d_0, d_1, \ldots, d_\nu\}$ where data in division categories above the maximum observed category are estimated using $E[Z_k]$. Alternatively, all data beyond $d_{\eta-1}$ can be accumulated into a single category in which case it has the form $\{d_0, d_1, \ldots, d_{\eta-1}, d_E = (d_\eta + d_{\eta+1} + \ldots + d_\nu)\}$. In the case where ν is finite we would estimate each d_i for $\eta \leq i \leq \nu$ using $E[Z_k]$. As $\nu \to \infty$ we would use $E[Z_\eta]$ as our estimator for the cumulative category. This yields the approximate log-likelihood functions:

$$l(\boldsymbol{\theta} \mid data) = \log\left(\frac{N_s!}{d_0!d_1!\ldots d_\nu!d_l!}\right) + d_l \log E[Z_{l,\nu}] + d_0 \log E[Z_0'] + \sum_{i=1}^{\nu} d_i \log E[Z_i] \tag{5.44}$$

in the first case and

$$l(\boldsymbol{\theta} \mid data) = \log\left(\frac{N_s!}{d_0!d_1!\ldots d_E!d_l!}\right) + d_l \log E[Z_{l,\nu}] + d_0 \log E[Z_0'] + \sum_{i=1}^{\eta-1} d_i \log E[Z_i] + d_E \log \sum_{i=\eta}^{\nu} E[Z_i] \tag{5.45}$$

in the second. For the second case we find that in the limit $\nu \to \infty$ we have

$$l(\boldsymbol{\theta} \mid data) = \log\left(\frac{N_s!}{d_0!d_1!\ldots d_e!d_l!}\right) + d_l \log E[Z_{l,\infty}] + d_0 \log E[Z_0'] + \sum_{i=1}^{\eta-1} d_i \log E[Z_i] + d_E \log E[Z_\eta] \tag{5.46}$$

Note that the use of accumulating all divisions above the maximum observed into one category is limited to predictions made by a model. We cannot treat actual data in this way since the data is in cell counts and we must normalize these to obtain the quantities of dye involved. By definition we do not know how many cells belong to each additional division category that the additional data may contain and therefore do not how to normalize this data.

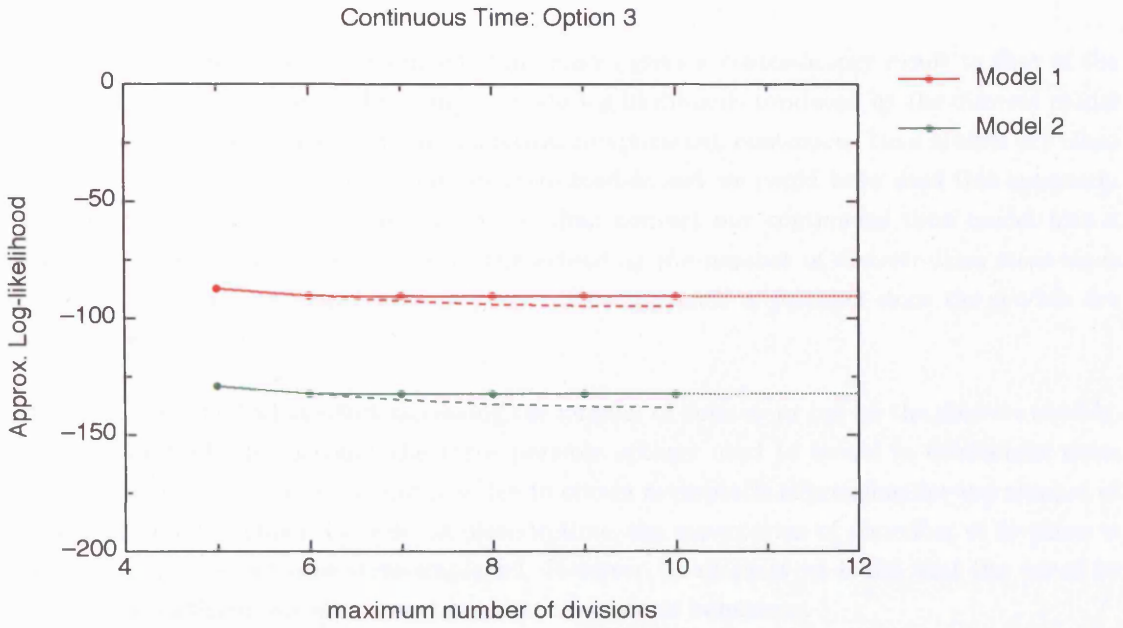

Figure 5.5: The effect of assuming the maximum number of divisions that have occurred is greater than that observed obtained by using the EM algorithm (option 3) on continuous time models 1 (red) and 2 (green). The likelihoods are those obtained using the data provided by Figure 6. H-2^d thymic stroma.(Hare et al., 1998). The results for both models are obtained by either accumulating all dye found in division categories greater than those observed (solid lines with asymptotes in dotted lines) or as separate categories (dashed lines). Results obtained by using the continuous time general model were identical to those of model 1.

The results of estimation using the EM algorithm also suggested that as the assumed maximum number of divisions increases so the approximate log-likelihoods decrease. This is true whether we treat the missing data separately by estimating with equation 5.44 (dashed lines) or cumulatively by estimating with either equations 5.45 or 5.45 (solid and dotted lines respectively). This suggests that, regardless of the model, the best fit is achieved for the observed data alone. In addition, EM estimation did not alter the relative positions of Model 1 and Model 2. Continuous time Model 1 has a better fit to the data than Model 2.

In summary, the overall result is that in the continuous case Model 1 fits the data better than Model 2 regardless of how the data is modelled. This contradicts the findings of the discrete modelling in which we used only the observed data which corresponds to our continuous option 1. At this point we therefore decided to investigate why this contradiction occurs. In doing so we expanded our discrete modelling to incorporate the approaches found in options 2 and 3.

5.3.6 Increasing the time steps for the discrete time model

In order to investigate why the continuous time model gives a contradictory result to that of the discrete time model, we examined the approximate log-likelihoods produced by the discrete model for increasing numbers of time steps. In numerical computation, continuous time models are often approximated by transforming them into discrete models and we could have used this approach. However in the interests of expediency, rather than convert our continuous time model into a discrete time approximation, we opted for the extending the number of discrete time steps since programs for the discrete model were in place. This approach is justified since the models are analogues of each other.

We therefore examined what effect increasing the number of time steps has on the discrete models. In doing so we took into account the three possible options used to model in continuous time. Note that in the discrete case it is not possible to obtain asymptotic expressions for the number of time steps at infinity. This is because, in discrete time, the expectation of a number of divisions is coupled to the number of time steps employed. However, in all cases we found that the use of 60 time steps was sufficient for all curves to exhibit asymptotic behaviour.

Using the experimental data provided Figure 6. H-2^d thymic stroma (Hare et al., 1998), the effects of increasing the number of time steps in the discrete models on the approximate log-likelihood are seen in figure 5.6. In addition to models 1 and 2 we also show the log-likelihoods produced by the discrete general model (UGM). In general the result indicates that the effect of increasing the number of time steps is to increase the likelihood for model 1 and decrease the likelihood of model 2. In all cases, the likelihood likelihood of model 1 becomes greater than model 2 as the number of time steps are increased. This would be in agreement with the result we obtained from the continuous model with model 1 providing the better fit.

However, the highest likelihood obtained is that produced by model 2 employing the minimum number of 6 time steps. In this case only option 1 applies since the data is not extended beyond the observed values. This indicates that the best fit to the data is found if we model the cells as only having undergone 5 divisions. Note also that, for any given model, the highest likelihoods were obtained using option 1, further supporting this conclusion. In addition under option 1 with 6, 7 and 8 time steps, the likelihood produced by the general discrete model is identical to model 2 option 1. The MLEs produced by the general model were identical or close to those of model 2 option 1 also (figure 5.7). This result indicates that for time steps 6, 7 and 8, model 2 option 1 provides the best fit and cannot be bettered by the general model.

We also point out the the lowest likelihoods obtained were those using option 2 and these are distinctly lower (figure 5.6 b). This suggests that modelling the cells as having undergone a greater

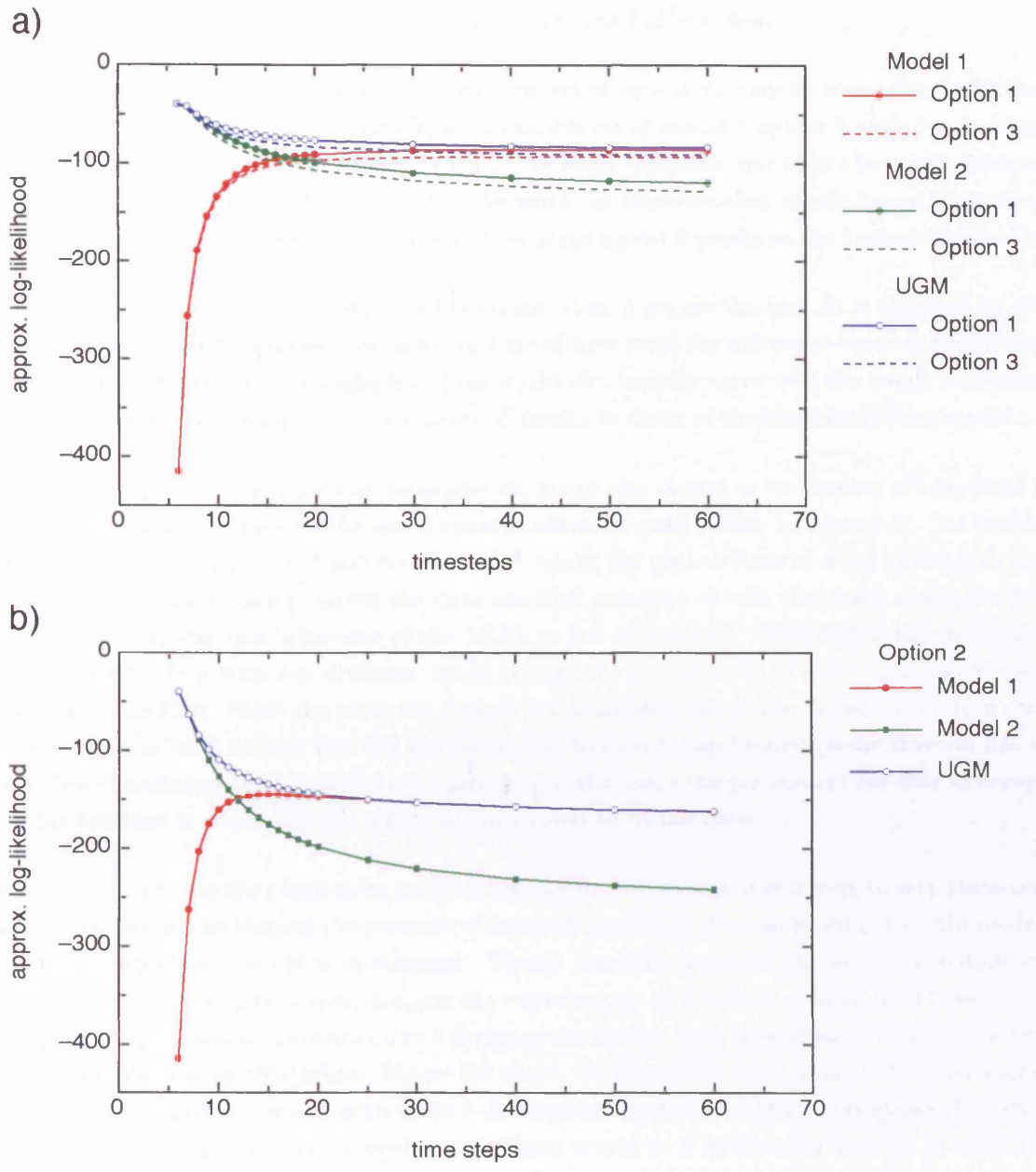

Figure 5.6: The effect of increasing the number of time steps on the approx. log-likelihood of discrete models 1, 2 and their general model (UGM) using a) options 1 and 3 and b) option 2. The likelihoods are those obtained using the data provided by Figure 6. H-2^d thymic stroma (Hare et al., 1998). When only 6 time steps were used, model 2 and the general model produced identical results. Under option 3 the results shown are derived using equation 5.44 and therefore assume that the additional division categories in the missing data are separate. The results for the alternative option 3 likelihood, where the missing data are accumulated into a single category (equation 5.45: result not shown), are similar to the results for option 3 with separate division categories.

number of divisions than those observed but with complete death of cells over and above the observed categories can be discounted as an adequate model of the data.

The similarity of option 1 and option 3 for low numbers of time steps may be an avenue for further investigation (figure 5.6 b). For example, is the likelihood of model 2 option 1 with 6 time steps significantly higher than model 2 option 3 with 7 time steps. However, our object here is to examine the difference between model 1 and model 2. We point out therefore that answering such questions will not alter the fact that for low numbers of time steps model 2 produces the highest likelihoods.

When the number of time steps employed is greater than 8 we see the best fit is supplied by the general model option 1. However, for large numbers of time steps the difference between the general model and model 1 option 1 is negligible. This would also broadly agree with the result that using the continuous general model produced identical results to those of the continuous time model 2.

As we would expect the values of our parameter estimates also change as the number of time steps is increased and as an example we discuss the results obtained using option 1 (figure 5.7). Noticeably, the probabilities of transition β and division γ fall, whilst the probabilities of doing nothing, δ_1 and δ_2, increase to close to unity. Given the data has high numbers of cells that have undergone few divisions, we consider this behaviour of the MLEs as not unexpected. With high numbers of time steps, fitting to data with few divisions would require the probabilities of doing nothing, at each time step to be high. Note also that the probability of division must also be kept low. However, there must be a limit to how low the probability of division γ can be since some division has to occur. We therefore suggest that the net result is that the space the parameters are able to occupy is so limited that it constrains the ability of the models to fit the data.

In the continuous case the parameters would also be under the constraint of having to keep transition and division low whilst keeping the prospect of doing nothing high. We can possibly view the models as being in some way caught in a dilemma. Firstly the data pressures the model to reduce the "tail" of the division distribution, keeping the expectations of divisions greater than 5 as low as possible. Secondly, where there are up to 5 divisions the models must also attempt to fit the pattern of the data. By way of illustration, Figure 5.8 shows the mean cell count from 100 simulations of continuous time models 1 and 2 with up to 9 divisions accounted for. This figure shows that given the MLEs the continuous models predict that there would be a substantial amount of cells that have divided more than 5 times. Given that these are the best fits to the data this suggest that the continuous models are limited in their capacity to reduce the "tail" of the division distribution whilst simultaneously fitting the categories where division has occurred.

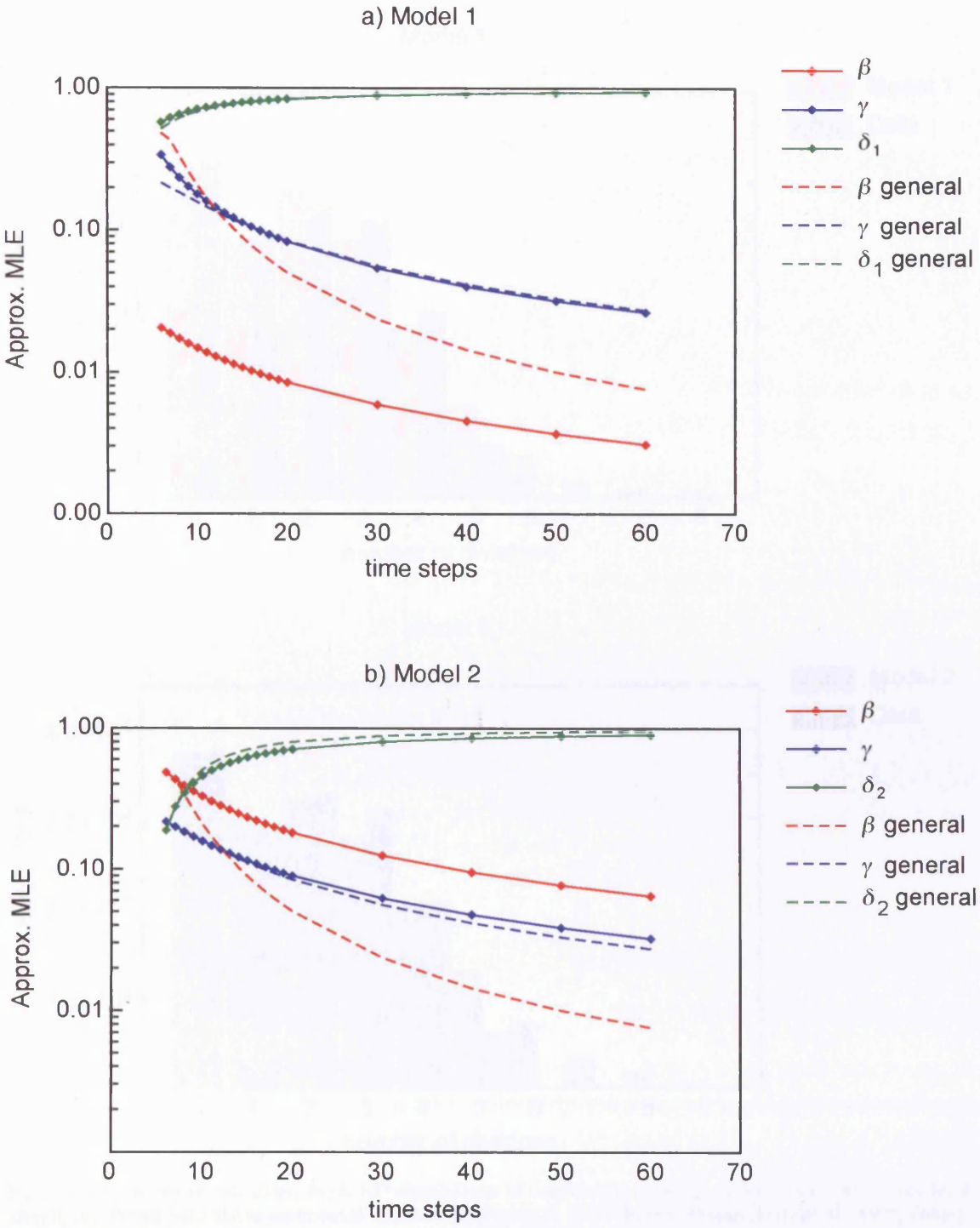

Figure 5.7: The effect of increasing the number of time steps on MLE values for models a) model 1 option 1 and b) model 2 option 1. The MLE values obtained under the general model are shown with both models for comparison.

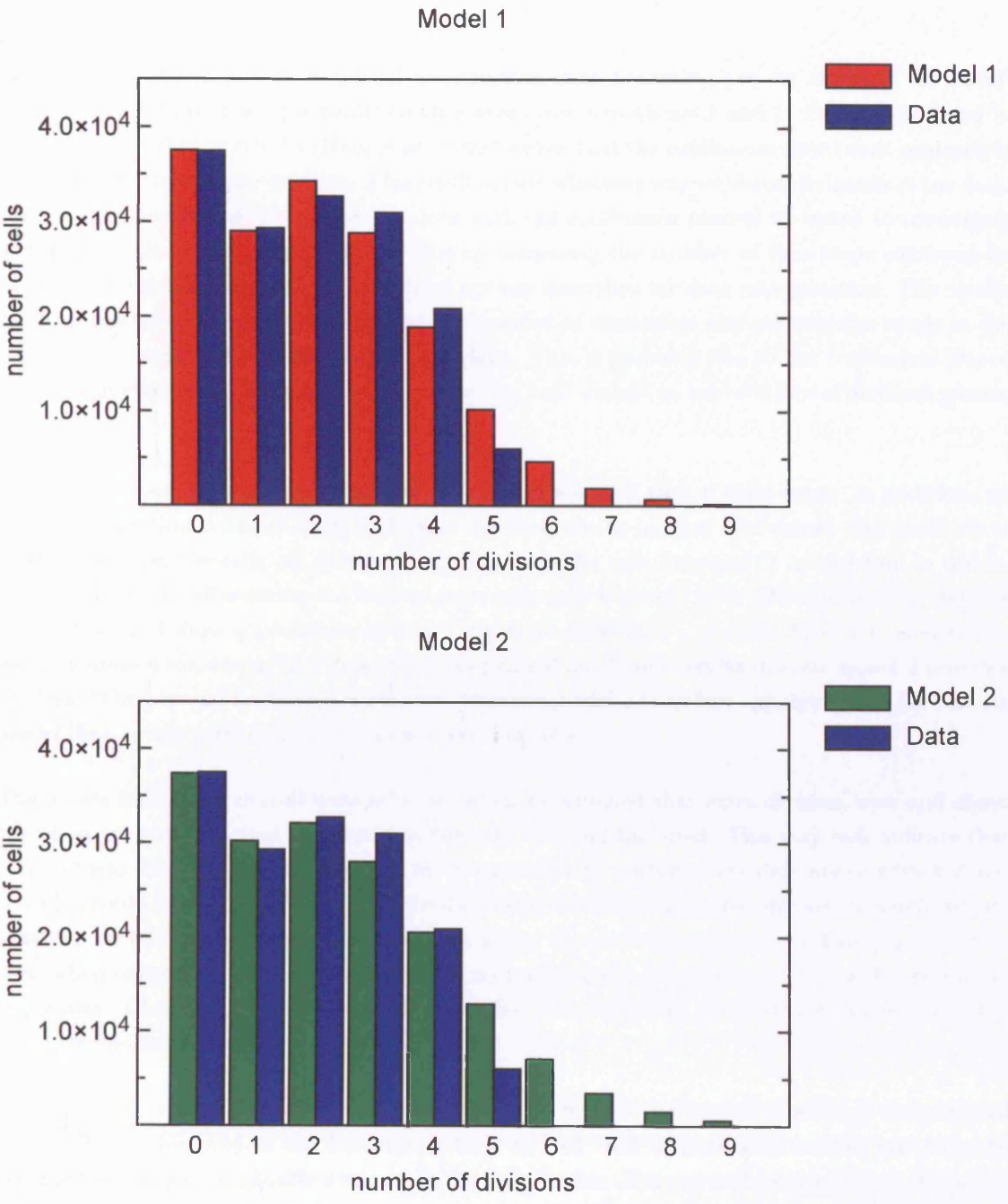

Figure 5.8: The mean cell count from 100 simulations of continuous time a) model 1(red) and b) model 2 (green) are shown with the experimental results from Figure 6. H-2^d thymic stroma (Hare et al., 1998) (blue). Up to 9 divisions are included in the simulated data. The parameter values used in the simulations were the MLEs derived from the experimental data; Model 1: $\beta = .05478$, $\lambda_1 = 1.17972$, $\gamma = 1.0$, $\lambda_2 = 0.51719$ and Model 2: $\beta = 1.0$, $\lambda_1 = 1.33832$, $\gamma = .32114$ and $\lambda_2 = 2.09153$.

5.4 Discussion

Here we have shown how to construct a continuous time branching process model of DP to SP transition. We have used this model to once again test hypotheses 1 and 2. Using data found in Figure 6. H-2^d thymic stroma (Hare et al., 1998) we see that the continuous time result contradicts that of the discrete time modelling. This result occurs whatever way we choose to interpret the data. Rather than continuing to analyse the data with the continuous models we opted to investigate why this contradiction occurs. We did this by increasing the number of time steps employed by the discrete models under the three distinct options described for data interpretation. The results of our investigation show that increasing the number of time steps also reverses the result in the discrete case regardless of how we view the data. This is probably due to the constraints placed upon the models through attempting to reduce the "tail" caused be expectations of divisions greater than 5.

Overall, the best fitting model remains the discrete model 2 with 6 time steps. In addition, we also point out that it seems likely that there is a limit on the number of divisions that could occur in the time that the cells are cultured. Mammalian cells take between 10 to 12 hours to divide. Conservatively, upon entering the culture some cells may be ready to divide immediately and this means that in 3 days a maximum of 6 or 7 divisions could have occurred. With our models this would require a maximum of 8 time steps. As pointed out in our results discrete model 2 provides the best fitting model for time steps 6 to 8. Discrete model 2 therefore appears be the best fitting model that agrees with these physiological expectations.

The lowest likelihoods overall were achieved when we assumed that extra division, over and above the observed numbers, had occurred but the cells involved had died. This may well indicate that this scenario did not occur. However, if more sophisticated models of the data are constructed and tested it would be advisable to examine the data using all three alternative options. It is unfortunate that the use of cumulative categories for data above the observable number of divisions cannot be used when examining data. The difficulty of normalizing the cell counts rendering this procedure impossible. Clearly, if normalization were possible then we would automatically know how many divisions had taken place.

In conclusion, we suggest that the continuous time models are poorer models of the analysed data. This is mirrored in the literature where we find that discrete time models are generally regarded as the most applicable when modelling cell division (Kimmel and Axelrod, 2002). However, in maximum likelihood the basic principle is that the result is dependent on the data and the continuous approach described here may therefore prove useful when used elsewhere.

Chapter 6

The Effect of Division and Selection

6.1 Introduction

The potential repertoire of TCR clonotypes derived through the random recombination of receptor gene segments is enormous. Estimates suggest that the number of possible rearrangements lies in the region of 10^{15} (Regner, 2001; Holler et al., 2003). However, many of these receptors will be unable to recognize anything at all and the thymocytes that bear them will either commit to further gene re-arrangement or apoptose. A proportion of those cells with TCR that can recognize self peptide presented on MHC will be auto-reactive and are therefore thought to undergo a process of negative selection in the thymus.

It has been suggested that negative selection is not 100% efficient at deleting potentially auto-reactive cells (Bouneaud et al., 2000). Mechanisms therefore exist that neutralize the ability of escapees to react to self, such as anergy and regulation (Seddon and Mason, 2000; Beissert et al., 2006). The cause of this inefficiency maybe related to the mechanism by which cells undergo negative selection and this is believed to involve multiple encounters with APCs (figure 1.2) (Sebzda et al., 1999). By encounter we mean that a thymocyte makes contact with an APC and "samples" the p/MHC presented by that cell. An encounter that results in a strong or high affinity interaction between the thymocyte and APC via TCR signalling induces apoptosis. If the interaction is weak the thymocyte survives and continues sampling. Such a process is inherently stochastic and therefore there always remains a finite possibility of escape.

Thus far we have used branching process models to analyse experimental data. This work produced evidence that suggests that negative selection possibly takes place while cells are dividing. According to Hare et al. (1998) this division is probably due to the effect of the cytokine IL-7 and therefore

independent of MHC interaction. Intuitively we would suspect that the addition of division to the sampling mechanism described above would only result in increasing the number of escapees. We therefore now investigate what would be the effect of a dividing and selecting process on the final repertoire of thymocytes. Here it is not necessary to use a staged model featuring transition from one phenotype to another. We merely assume that a population of thymocytes exists and this is subject to negative selection. Subsequently, we use a simple branching process as our model. By way of comparison we also model the effect of non-dividing selection. Principally, our results indicate that division during negative selection can increase the number of clonotypic specificities whilst having a very small effect on the rate of auto-reactivity.

6.2 Methods

6.2.1 The Process We Are Modelling

The process of negative selection can be likened to a game of russian roulette. The bullets in this game are TCR signals, received during an encounter with an APC, that are sufficiently strong enough to induce apoptosis. Each encounter can therefore be considered to be akin to squeezing the trigger. At any given encounter, we can therefore view a thymocyte's probability of death (α) as the probability that it will receive the killer signal. If it does not receive the death signal it continues to the next encounter; pulling the trigger once more.

Here, we augment this scenario by including division which occurs with probability γ. However, in line with Hare et al. (1998) we do not assume that division is related to thymocyte APC interaction. We also assume that daughter cells produced by division continue experiencing encounters following their division event. Division can therefore be thought of as a background process on to which negative selection is overlaid. One final possibility remains, if a cell neither divides or dies it survives an encounter with probability δ.

6.2.2 The Branching Process Model

Following on from the above, we therefore model a thymocyte as having three fixed properties: its probabilities of death α, division γ and stasis δ at each encounter k (figure 2.6). We assume each of these events to be mutually exclusive such that $\alpha + \gamma + \delta = 1$. We assume division to be binary ie. a cell divides into 2 daughter cells. Also, all descendants of a cell inherit its specific α, γ and δ values.

Given these assumptions we can model the process of negative selection as a simple branching process with generating function:

$$G(s) = \alpha + \delta s + \gamma s^2 \tag{6.1}$$

Often the time step employed when using this type of model is set to that of the cell cycle (Jagers, 1975). Here we set the time step k to be that of an encounter. It is thought that the average duration of an encounter is between 5 and 10 minutes (Muller and Bonhoeffer, 2003). This presents us with the possible dilemma in that our process appears to make the non-biological assumption that a cell divides within this time span. However we have found that, provided division events are rare ie. $\gamma << 1$ (as we assume here), our model gives qualitatively acceptable results.

To recap, the mean or expected number of individuals at time step $E[Z_k]$ is given by raising the differential of $G(s)$ with respect to s at $s = 1$ to the power k ie. $[G^{'}(1)]^k$ (Harris, 1963; Jagers, 1975). For G(s) we therefore obtain,

$$E[Z_k] = (\delta + 2\gamma)^k \tag{6.2}$$

We can also compute the probability that the process will be extinct at encounter k and this is given by $G_k(0)$ (Harris, 1963; Jagers, 1975). We define P_k as the probability that a branching process is not extinct at encounter k, namely $P_k = P(Z_k > 0)$. Given that $G_k(0)$ is our probability of extinction at encounter k then clearly the probability of being alive is,

$$P_k = 1 - G_k(0) \tag{6.3}$$

We can also compute the ultimate probability that the branching process will become extinct ie. $\lim_{k \to \infty} G_k(0)$ by finding the smallest solution of $s = G(s)$ (Harris, 1963; Jagers, 1975);

$$s = G(s) = \alpha + \delta s + \gamma s^2 \tag{6.4}$$

which after rearrangement becomes the quadratic

$$\gamma s^2 + s(\delta - 1) + \alpha = 0 \tag{6.5}$$

when the usual methods for solving quadratics are applied the roots are

$$s = \lim_{k \to \infty} G_k(0) = \frac{\alpha}{\gamma} \text{ or } 1 \tag{6.6}$$

Which of the roots is the smallest depends on the relative values of α and γ. If $\alpha = \gamma$ the roots of equation 6.5 are both equal to 1. When $\alpha > \gamma$ we see that $\alpha/\gamma > 1$, so the smallest root of equation 6.5 is $s = 1$. When $\alpha < \gamma$ we find $\alpha/\gamma < 1$ and therefore supplies the smallest root. In biological terms, if the probability of death is greater than or equal to the probability of division then the ultimate probability of extinction is 1.

In additon, two trivial special cases also occur, a) when $\alpha = 0$ or 1, since when $\alpha = 0$ no cell will ever die and there is no probability of extinction and b) when $\alpha = 1$ all cells will die at time step 1 and therefore the probability of extinction is always 1.

From equations 6.3 and 6.6 we see that either

$$\lim_{k \to \infty} P_k = 1 - \lim_{k \to \infty} G_k(0) = \left\{1 - \frac{\alpha}{\gamma}\right\} \text{ or } 0 \tag{6.7}$$

Barring the special cases and following the logic applied above, here we see that the when $\alpha = \gamma$ the LHS of equation 6.7 will equate to zero. Also when $\alpha > \gamma$ the smallest root of equation 6.5 is $s = 1$ so the LHS of equation 6.7 also equates to zero. So only when $\alpha < \gamma$ does the branching ultimately survive with some finite probability.

In addition to the above we note that, the probability of the process being extinct at time step k ie. $G_k(0)$ is a sum of terms in α. At time step $k + j$ (where j is a number of additional time steps) we therefore have

$$G_{k+j}(0) = G_k(0) + \omega \tag{6.8}$$

where ω represents additional positive terms in α. As k tends to infinity we would expect ω to tend to zero since

$$\lim_{(k+j)\to\infty} G_{k+j}(0) = \lim_{k\to\infty} G_k(0) \tag{6.9}$$

Therefore, as k tends to infinity, $G_k(0)$ monotonically increases towards its assymptotic limit.

6.2.3 Contour Plots

It is a matter of debate as to how many encounters a thymocyte undergoes during its stay in the thymus (Scollay and Godfrey, 1995). Indeed, there is evidence to suggest that the duration of a thymocyte's stay in the thymic medulla varies in the region of 5-14 days (Gabor et al., 1997). Given an encounter duration of 5 minutes, a thymocyte may therefore undergo roughly somewhere between 1400 to 4000 encounters. In order to visualize the effect of negative selection on P_k over such large ranges of possible k values, we view fixed levels of P_k as contours in α and k parameter space (figures 6.2, 6.4 and 6.5). This has considerable advantage over some previous attempts to model negative selection (Muller and Bonhoeffer, 2003) since no *a priori* assumptions are made as to the exact number of encounters.

Plotting Method

A contour plot corresponding to a probability η of surviving selection is given by solving the equation

$$P_k(\alpha) - \eta = 0$$

for α. Here, a reasonable compromise between speed and accuracy can be achieved by setting the computational precision (ϵ) in finding the solution to $\epsilon = 1 \times 10^{-5}$.

6.2.4 The Effect on a Distribution of Cells

The methods above allow us to examine the effect of non-dividing and dividing selection on individual cells for a given parameter regime. However, it is not clear what the cumulative effect of our modelling would be on an entire distribution of post-positively selected cells. We therefore set out to examine the effect of our modelling using standard Bayesian methods. In this approach we assume an initial post-positive selection distribution of cells on α, referred to as the prior distribution. The distribution of remaining cells after negative selection is therefore referred to as the posterior distribution. Since neither the prior nor posterior distributions are known we have chosen to model the prior distribution using the log-normal density function (see below).

The choice of a log-normal prior distribution can be justified in that it fits 2 basic criteria: firstly, the log-normal density function is always 0 at $\alpha = 0$; thus mimicking the effect of positive selection in which all non-reacting clones are removed from the repertoire. Secondly, we hoped initially to be able to examine the effect of changing the variance of our prior distribution on the posterior distribution of cells, and therefore considered it an advantage that the log-normal variance is independent of its mean or median.

Examination of the behaviour of the log-normal density function reveals that with a fixed mean changing the variance has a dramatic effect on the shape of the distribution (results not shown). On increasing the variance, this effect manifests itself by increasing the distribution's positive skew. This type of behaviour may be typical of many continuous positively valued distributions with mean/median close to zero.

Whatever prior distribution is adopted, the end result here is a posterior distribution of post-negatively selected cells. A caveat to all this is that a distribution of thymocytes can be viewed in 2 ways: a distribution of clones or a distribution of individual thymocytes. We therefore examined both these cases as follows.

Clone distributions

First we define $P(\alpha)$ (not to be confused with $P_k(\alpha)$) as the prior distribution density function that provides the probability that a clone picked at random has a specific α value. As mentioned above biologically we do not know what form $P(\alpha)$ takes but have chosen to model this using the lognormal density function,

$$P(\alpha) = \frac{1}{\delta\alpha\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{\ln\alpha - \mu}{\sigma}\right)\right\} \tag{6.10}$$

where $\ln\alpha$ is normally distributed with mean $\mu = \mathrm{E}[\ln\alpha]$ and variance $\sigma^2 = \mathrm{Var}[\ln\alpha]$. In α this distribution has mean (M) and variance (S),

$$\mathrm{E}[\alpha] = \mathrm{M} = \exp\left\{\mu + \frac{1}{2}\sigma^2\right\} \tag{6.11}$$

$$\mathrm{Var}[\alpha] = \mathrm{S} = \exp\left\{2\left(\mu + \frac{1}{2}\sigma^2\right)\right\}\left[\exp\left\{\sigma^2\right\} - 1\right] \tag{6.12}$$

Re-arrangement of 6.11 and 6.12 and subsequent substitution into 6.10 yield,

$$P(\alpha) = \frac{1}{\alpha\sqrt{2\pi}\sqrt{\ln\theta}} exp\left\{\frac{\left(\frac{1}{2}\ln\theta + \ln\alpha - \ln\mathrm{M}\right)^2}{2\ln\theta}\right\} \tag{6.13}$$

where,

$$\theta = 1 + \frac{S}{M^2}$$

thus we obtain our prior distribution function with mean M and variance S.

Using Bayes theorem our posterior distribution is defined as the probability that a clone picked at random from the clones that survive negative selection has a particular α value ($P(\alpha \mid Clone_{surv})$). This is obtained by multiplying our initial density function $P(\alpha)$ by the probability of surviving negative selection $P_k(\alpha)$, giving the probability of surviving negative selection conditioned on our prior distribution. This is then normalised using the integral of this product as a denominator. Thus our posterior distribution is provided by the conditional probability density function

$$P(\alpha \mid Clone_{surv}) = \frac{P(\alpha)P_k(\alpha)}{\int_0^1 P(\alpha)P_k(\alpha)d\alpha} \tag{6.14}$$

Note that since $\alpha + \delta + \gamma = 1$ we define

$$P(\alpha) = 0 \quad \text{for} \quad \alpha > 1 - \gamma$$

Choosing a Prior Distribution Mean

Our choice of a prior distribution mean $M = .002$ was roughly estimated by observation of the contour plots. This was based on the generally held view that negative selection deletes somewhere between 67 and 90% of positively selected thymocytes (Mason, 1998). This would loosely correspond to the 30 and 10% survival contours. In addition, using the lower limit of 5 days duration in the thymic medulla and a rough estimate of 1 encounter every 5 to 10 minutes we would expect that the number of encounters $k > 1000$. This would place a mean α somewhere in the range of .0015 to .0025 (figure 6.4) for the lowest k estimate. Thus taking the mean of .002 provides a very rough estimate of a prior distribution mean.

Thymocyte Distributions

Here our prior distribution of thymocytes $F(\alpha)$ is defined as the probability that a thymocyte picked at random has a specific α value, irrespective of clonal specificity. In order to obtain a posterior distribution of thymocytes we use the expected number ($E[Z_k]$) of thymocytes for a branching process at the kth encounter. Similarly to equation 6.14 above we therefore derive from Bayes theorem the conditional density function,

$$P(\alpha \mid Cell_{surv}) = \frac{F(\alpha)E[Z_k]}{\int_0^1 F(\alpha)E[Z_k]d\alpha} \tag{6.15}$$

In words, $P(\alpha \mid Cell_{surv})$ is defined as the probability that a thymocyte picked at random from the thymocytes that survive negative selection has a particular α value. We assume, for the moment, that in our prior distribution each clone present has an initial population size of 1 thymocyte so

that $F(\alpha) = P(\alpha)$. This is based on the assumption that the large number of possible TCRs generated by gene rearrangement would mean that the chance of any 2 post-positive thymocytes bearing identical TCR would be very small.

Thymocyte Numbers

If the number of thymocytes in an initial population is large then the numerator of equation 6.15 can be used to derive the expected number of thymocytes at time step k given our initial starting distribution $F(\alpha)$ as in,

$$E[Z_k \mid F(\alpha)] = E[Z_k]F(\alpha) \tag{6.16}$$

Terminology

In simple terms the distributions produced by $P(\alpha \mid Clone_{surv})$ and $P(\alpha \mid Cell_{surv})$ can be described as frequency distributions. This is in contrast to the curves for $E[Z_k \mid F(\alpha)]$ which refers to the final number of thymocytes produced. In order to simplify further reference to these distributions, we therefore refer to the former as frequency distributions and the latter as number distributions.

6.3 Results Relating to Individual Cells

6.3.1 Non-Dividing Selection

The case where selection occurs in the absence of division can be modelled using our generating function by setting $\gamma = 0$. In this case, we find $\delta = 1 - \alpha$. Substituting into equation 6.3 and subsequent simplification reveals that at kth encounter

$$P_k = (1 - \alpha)^k \tag{6.17}$$

We note that this is the lower bound of P_k for our branching process as P_k is a sum of terms in α, δ and γ (see equation 2.4 for an example) and setting $\gamma = 0$ effectively removes terms that include γ.

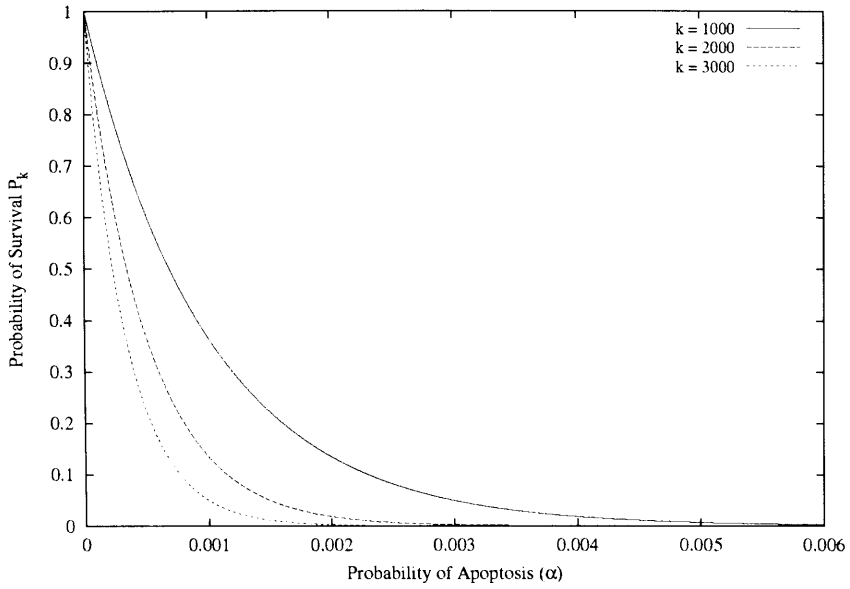

Figure 6.1: In the absence of division the probability of surviving negative selection $P_k = (1-\alpha)^k$ decreases monotonically with increasing the number of encounters (k) for $0 < \alpha < 1$. Here, in aide of clarity we truncate the x-axis and present 3 curves indicating the value of P_k over the range $0 < \alpha \leq 0.006$ for: $k = 1000$ (solid); $k = 2000$ (dashed) and $k = 4000$ (dotted). Taking any particular value of α in the displayed range we see that increasing k results in a reduction of survival probability.

Also, for a thymocyte with with probabilities $\alpha : \{0 < \alpha < 1\}$, $\delta = 1-\alpha$, and $\gamma = 0$ the probability P_k decreases monotonically with increasing k (figure 6.1). However, since $\gamma < \alpha$ we see that from equations 6.3 through 6.9 and associated commentary it is also true that only in the limit

$$\lim_{k\to\infty} P_k = \begin{cases} 0 & 0<\alpha\leq 1 \\ 1 & \alpha=0 \end{cases} \tag{6.18}$$

from this and equation 6.17 we see that for $0 \leq \alpha \leq 1$ with any $k < \infty$ our $P_k > 0$.

The sensitivity of P_k to k can be examined using the contour plots (figures 6.2, 6.4 and 6.5). In the case of non-dividing selection, this sensitivity increases with increasing α, as indicated by the relative vertical spacing of contours in figure 6.2. Thus we see that for relatively higher values of α the contours occupy a relatively small range of k values; eg. for $\alpha = .0015$ and $P_k = 10$ to 90%, $k :\sim 50$ to ~ 1600. When we compare this finding to the lower values of α we see that k makes little

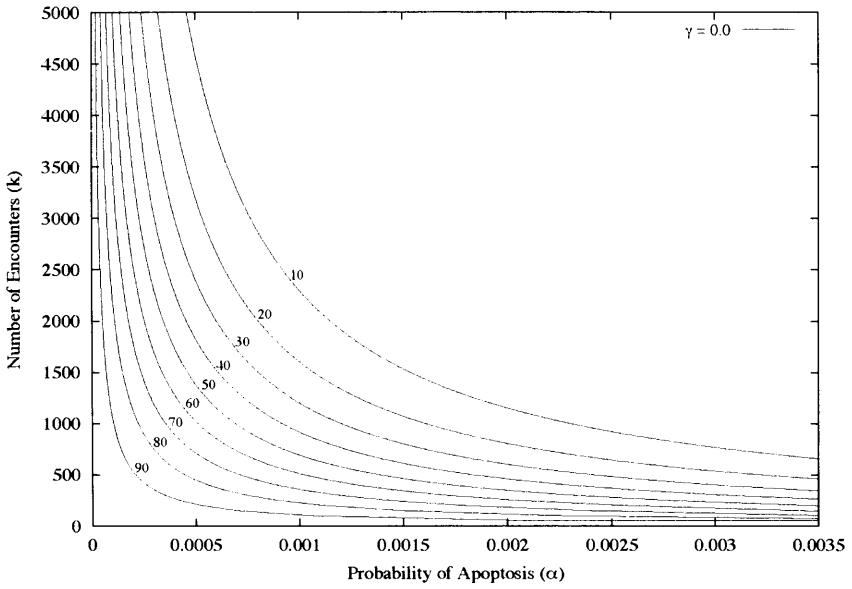

Figure 6.2: In the absence of division, the sensitivity of the probability of survival (P_k) to the number of encounters (k) increases with increasing α. The contours show P_k = 10, 20, 30,...,90% probability of survival in α and k parameter space.

difference to the probability of survival; eg. for $\alpha = .0005$ and $P_k = 10$ to 90%, $k :\sim 200$ to ~ 4700.

6.3.2 Division During Negative Selection

As with the non-dividing case, from equations 6.3 through to 6.9 and associated text we see that with increasing k our P_k monotonically decreases for all α in the range $0 < \alpha < 1$. However, equation 6.7 shows that, in the limit $k \to \infty$, $P_k = 1 - \alpha/\gamma$. If we view P_k as a function of α and provided γ is a constant this limit is linear with gradient $-1/\gamma$ intersecting the y-axis at 1 and the x-axis at γ (figure 6.3). This means that for clones with a γ value below α there is always some probability of survival even if $k = \infty$.

Thus far it seems that we can with certainty predict the differences between selection with division and without as $k \to \infty$. However, the maximum value for k is thought to be in the region of c.a. 4000 encounters (Muller and Bonhoeffer, 2003). What then is the difference in the effect of division on selection for relatively low values of k?

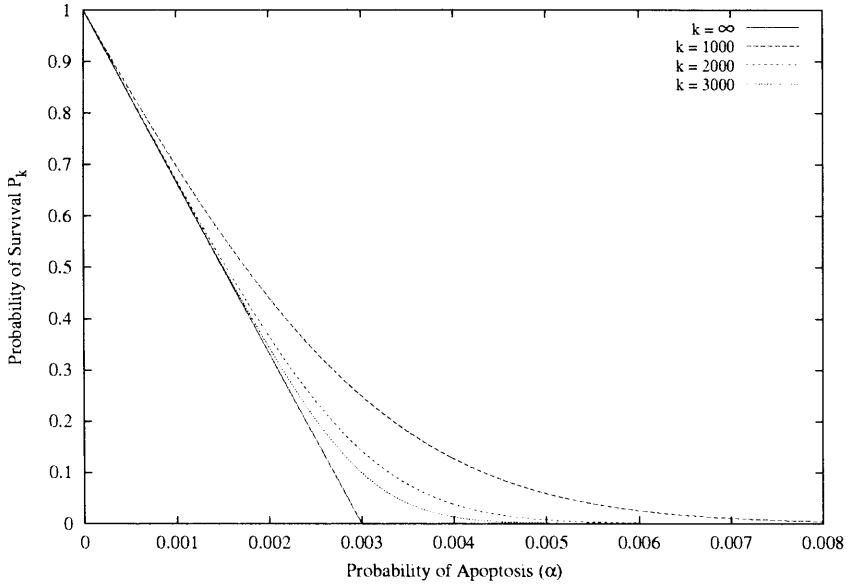

Figure 6.3: In the presence of division the probability of surviving negative selection $P_k = 1 - G_k(0)$ (equation 6.3) decreases monotonically in the range $0 < \alpha < 1$ towards the $\text{limit}_{k \to \infty} P_k = 1 - \alpha/\gamma$ (solid). For clarity the x-axis is truncated and we show 3 curves representing P_k at $k = 1000; k = 2000$ and $k = 3000$. In all cases the probability of division $\gamma = 0.003$. Note: Plotting the curves can be achieved analytically in either Mathematica or Maple by setting up $G_k(s)$ as a nested function with a subsequent call to the function at $G_k(0)$.

To answer to this question we turn to figures 6.4 and 6.5, where we see that for k in the range 0 to 5000 encounters and in the presence of division, the increase in sensitivity to k with increasing α follows the same qualitative pattern as in its absence. Noticably, however, division has the effect of making the contours become relatively more laterally separated at higher k values. This indicates a broadening of the range of α over which P_k has low sensitivity to k. Indeed, an alternative way of viewing the same effect is to observe P_k fixed whilst varying our division parameter γ (figure 6.5). Here increasing the γ has the effect of resetting the value of α required to produce equivalent levels of P_k. This effect is more exaggerated for lower values of P_k, as indicated by the wider lateral spacing of contours in figure 6.5a when compared to figure 6.5b or 6.5c.

When comparison is made to non-dividing selection, these results indicate that for physiologically relevant numbers of encounters, division during selection not only reduces the sensitivity of the probability of survival P_k to the number of encounters k but also broadens the range of α over which it has a low sensitivity to the number of encounters. Furthermore, increasing the amount of

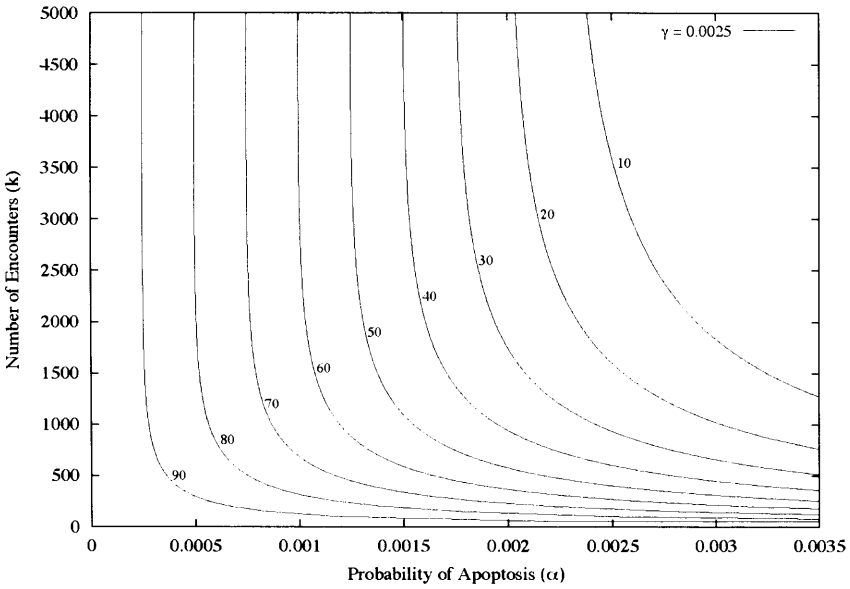

Figure 6.4: In the presence of division the sensitivity of P_k to k increases with increasing α. Contours show P_k = 10, 20, 30,...,90% probability of survival when the probability of division $\gamma = 0.0025$ in α and k parameter space.

division (increase γ) acts to increase these 2 effects and the impact of this is more strongly felt at low α. In biological terms division increases the chances of survival of a clone and the greater the division the more likely a clone is to survive.

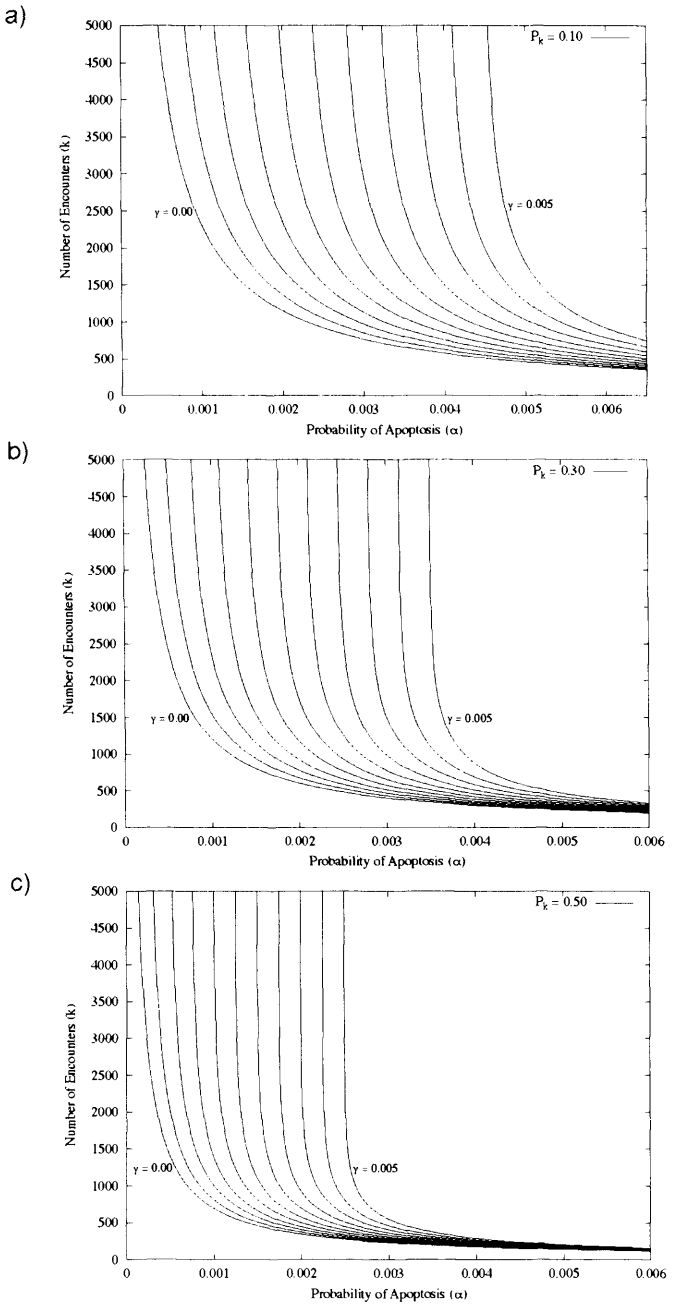

Figure 6.5: The effect of varying the probability of division γ on P_k. The contours show P_k = a) 0.10 , b) 0.30 and c) 0.50 probabilities of survival in α and k parameter space for values of γ = 0 to .005 in .0005 increments. The lateral displacement of the contours is larger for lower P_k.

6.4 Results Relating to Distributions of Cells

6.4.1 Frequency of Non-Dividing Clones

Given a log-normally distributed prior distribution of clones that do not undergo division during negative selection, our results indicate that negative selection would produce final frequency distributions with a distinct sharply peaked positive skew (figure 6.6). In biological terms this indicates that only the clones in our initial distribution with very low probability of death (α) survive. In agreement with our previous results this effect is more exaggerated when the number of encounters is increased.

6.4.2 Frequency of Dividing Clones

Given the same starting distribution as non-dividing clones, the effect of division would appear to increase the breadth of the final distribution on α (figure 6.7). The biological interpretation of this result is that division results in greater survival of relatively high α clones. This result agrees with the results in figures 6.4 and 6.5 where we saw that relatively higher values of α were required in order to produce equivalent levels of deletion to those seen in non-dividing selection.

In comparison to the results for non-dividing clones, where increasing k leads towards a more sharply peaked distribution, it is noticeable that the posterior frequency distribution of clones heads towards a limit as we increase k. In particular, there is little difference in the posterior distributions of dividing clones when we compare the curves for 3000 and 4000 encounters. This is a direct consequence of the increased insensitivity to k incurred by division.

6.4.3 Numbers of Non-Dividing Thymocytes

It is obvious that the effect of negative selection without division on the number of thymocytes to reduce their total numbers (figure 6.8). This reduction in numbers is, in agreement with our previous results, sensitive to k. This is clearly seen in figure 6.8b, where the entire initial distribution is nearly eliminated at $k = 4000$

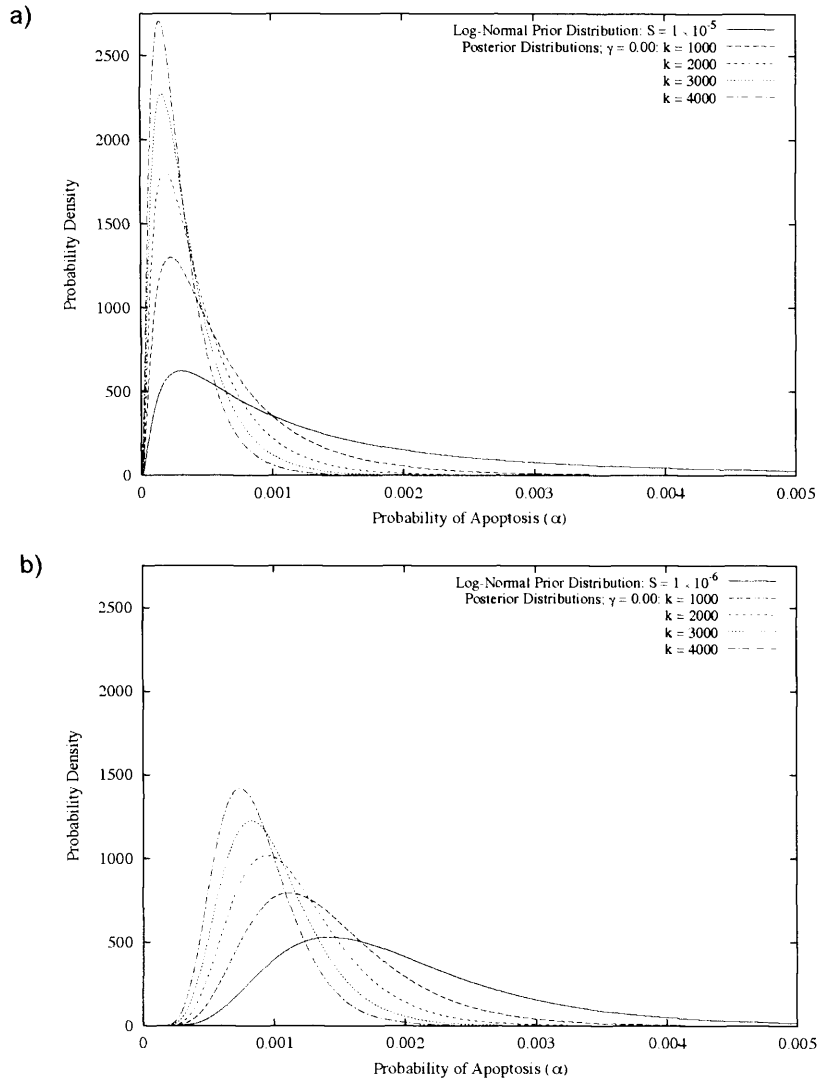

Figure 6.6: The effect of negative selection in the absence of division on a log-normal prior distribution of clones (solid) with mean 0.002 and variance a) 1×10^{-5} and b) 1×10^{-6}. The posterior distribution curves are the result of plotting equation 6.14 with $\gamma = 0$ and show the effect at $k = 1000$, 2000, 3000 and 4000 encounters. Clearly the distributions become more positively skewed with increasing k.

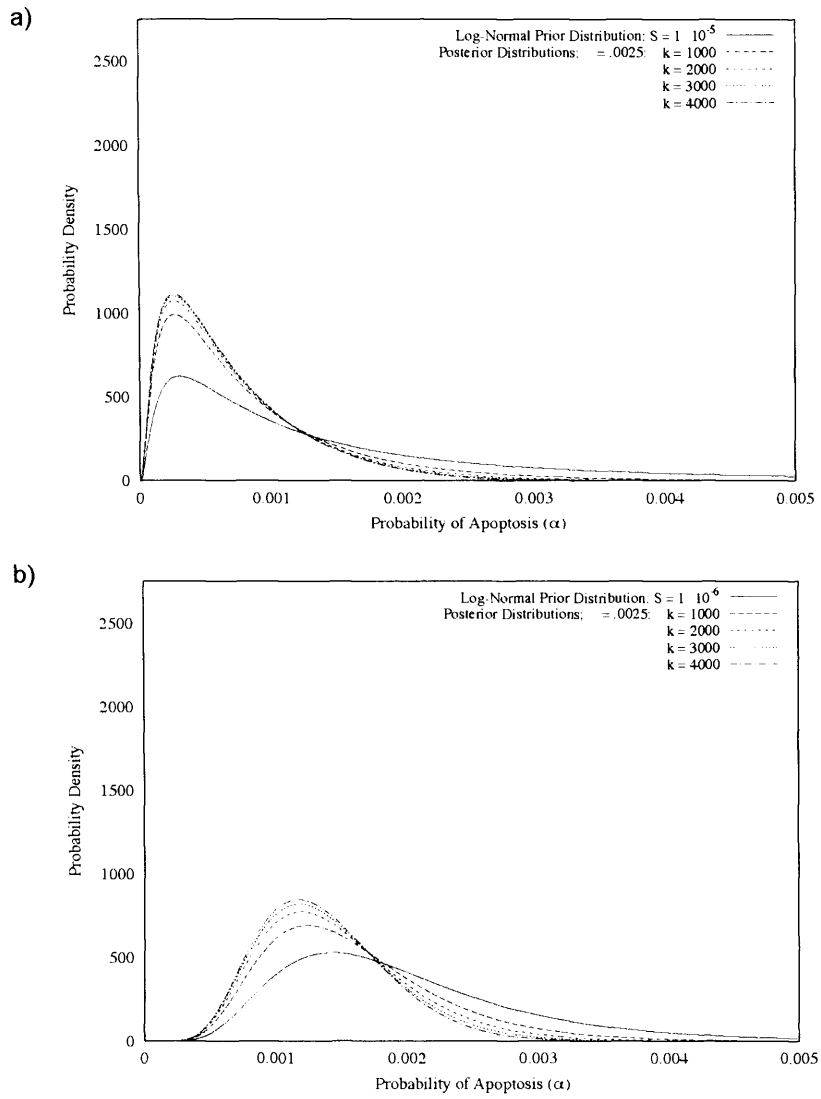

Figure 6.7: The effect of negative selection with division given a log-normal prior distribution of clones (solid) with mean 0.002 and variance a) 1×10^{-5} and b) 1×10^{-6}. The effect is shown for 4 values of $k = 1000$, 2000, 3000 and 4000 encounters with the probability of division $\gamma = 0.0025$. The posterior distributions are positively skewed and head to a limit (not shown) similar to the curve at $k = 4000$.

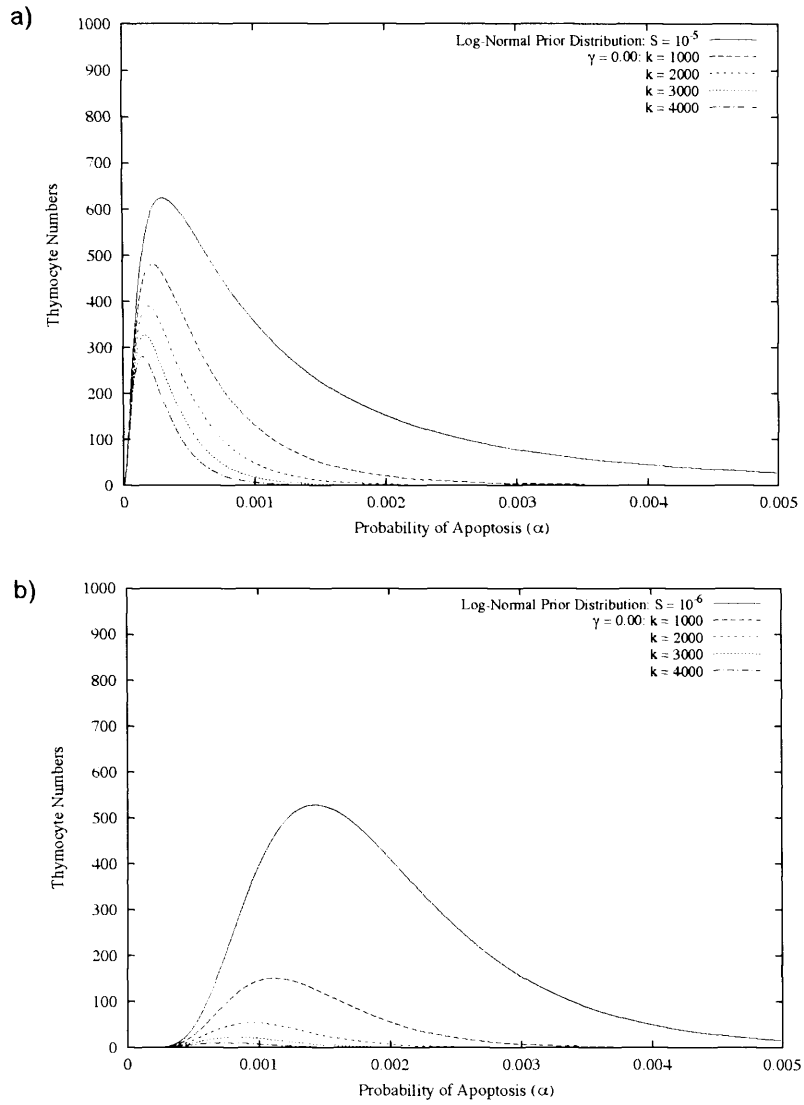

Figure 6.8: The effect of negative selection on a log-normal prior distribution of non-dividing thymocytes with mean $M = .002$ and variance a) $S = 10^{-6}$ and b) $S = 10^{-5}$. The solid curve represents the prior distribution of cells whilst the remaining curves show what happens as k is increased.

6.4.4 Numbers of Dividing Thymocytes

In contrast to the above, division during selection has a markedly different effect on thymocyte numbers (figures 6.9 and 6.10). Regardless of k, all the number distribution curves, including the prior distribution curve, have a common point of intersection (figures 6.9b and 6.10b). This intersection occurs at $\alpha = \gamma$. This is expected because substitution of $\delta = 1 - \alpha - \gamma$ into equation 6.2 at $\alpha = \gamma$ yields

$$E[Z_k] = (1 - \alpha - \gamma + 2\gamma)^k = (1 - \gamma - \gamma + 2\gamma)^k = 1 \tag{6.19}$$

Namely the expected number of cells $E[Z_k] = 1$ at $\alpha = \gamma$ regardless of the value of k. Note that the intersection of these curves creates a threshold about which the effect of dividing selection differs. Thymocytes with $\alpha > \gamma$ are reduced in numbers whilst those with $\alpha < \gamma$ are increased in number. This would mean that following selection clones with relatively high α values would have low copy numbers.

6.4.5 Frequency of Non-Dividing Thymocytes

Our assumption that each clone in the initial distribution ($F(\alpha)$) is present in the form of a unique thymocyte means that the frequency distribution curves for thymocytes undergoing non-dividing negative selection are identical to those obtained for non-dividing clones (figure 6.11 dotted lines cf figure 6.6). Since, without division, each thymocyte is the sole representative of a clone and remains so until it dies.

6.4.6 Frequency of Dividing Thymocytes

We have seen that division can have the effect of substantially altering the survival of a clone in comparison to non-dividing selection and this is reflected in the posterior frequency distribution of clones (figure 6.7). However, it paradoxically has little effect on the posterior frequency distribution of thymocytes (figure 6.11). Even when we raise the level of γ to 0.1 divisions per encounter the effect is negligible as indicated by the small difference between the zero division reference curves (dotted lines) and their corresponding dividing selection curves. This result is expected because,

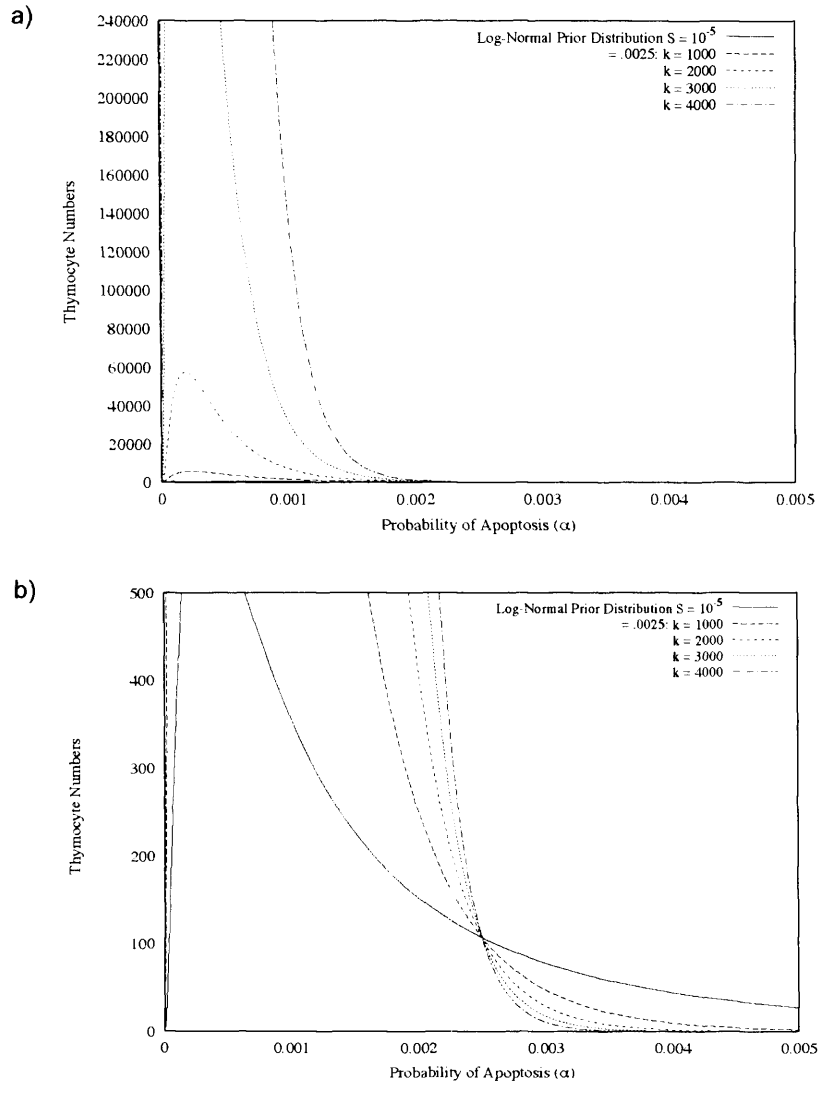

Figure 6.9: The effect on thymocyte numbers of negative selection acting on dividing cells with a log-normal prior distribution with mean $M = .002$ and variance $S = 10^{-5}$. Two views of the same plot are shown with different maximums for the y-axis a) max y = 240000 and b) max y = 500.

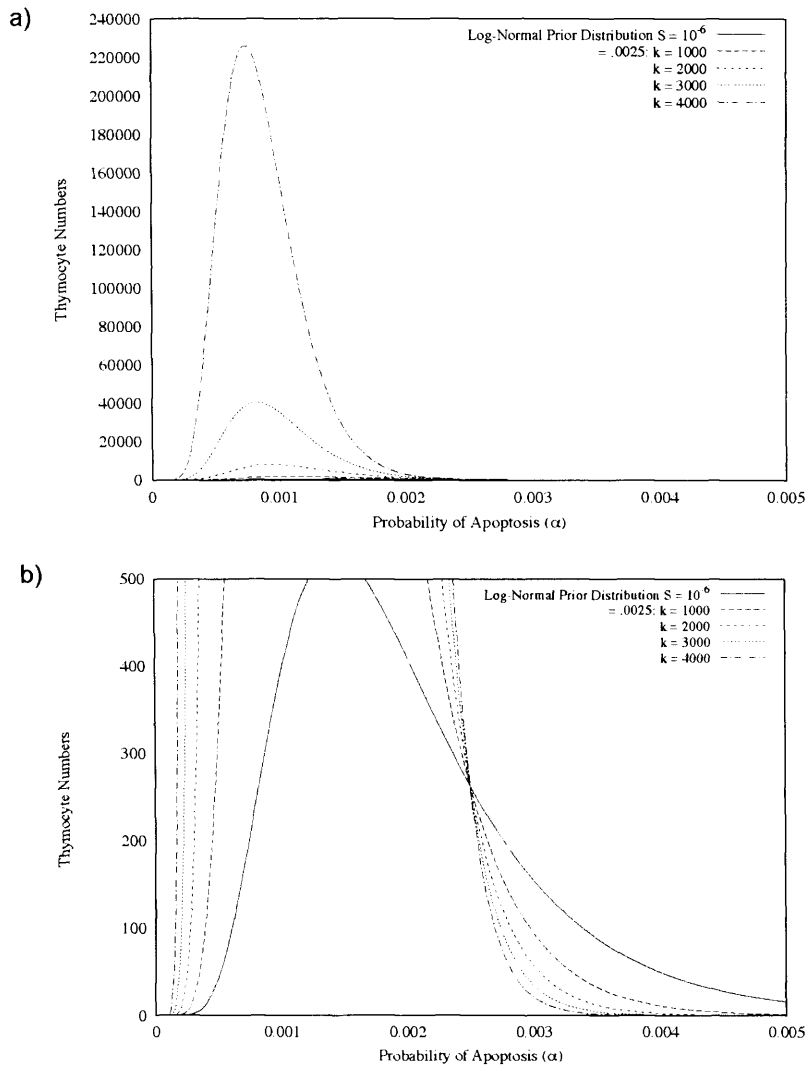

Figure 6.10: The effect on thymocyte numbers of negative selection acting on dividing cells with a log-normal prior distribution with mean $M = .002$ and variance $S = 10^{-6}$. Two views of the same plot are shown with different maximums for the y-axis a) max y = 240000 and b) max y = 500.

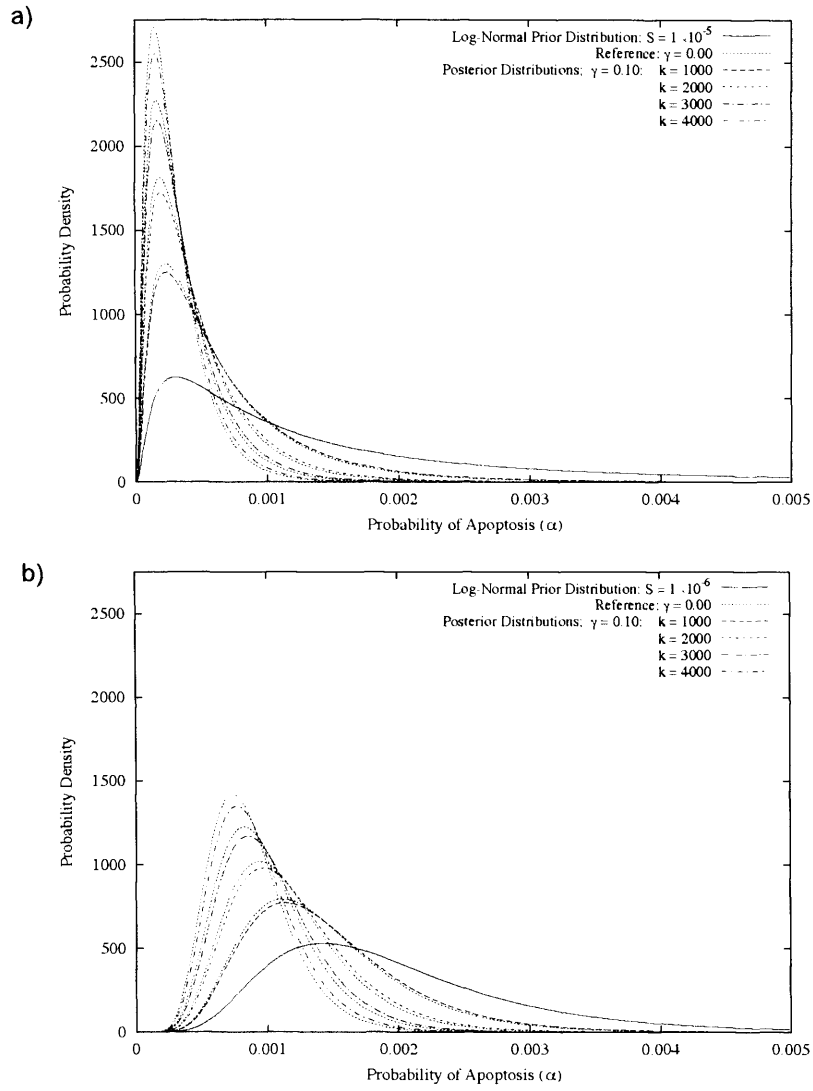

Figure 6.11: The effect of division and selection on a log-normal prior distribution of thymocytes (solid) with mean 2^{-3} and variance a) $S = 10^{-6}$ and b) $S = 10^{-5}$. Each division curve is accompanied by a reference curve for $k = 1000$, 2000, 3000 and 4000 encounters. Here $\gamma = 0.10$ and even at this relatively high level the relative difference in the curves for division and non-division (dotted reference) is negligible.

$$\lim_{\gamma \to 0} E[\alpha] = \lim_{\gamma \to 0} (\delta - 2\gamma)^k = (1-\alpha)^k \tag{6.20}$$

From the biological perspective, we can view this result as being in agreement with the results for numbers of thymocytes. This is because the higher α clones have low copy numbers (see above) and this would skew the distribution towards the low α clones.

6.5 Discussion

6.5.1 Improved Visualisation Through Contour Plots

It is self-evident that all thymocytes that survive positive selection have, to some extent, the capacity to detect self. At the very least, this self-awareness manifests itself as a basic affinity for MHC regardless of any specific epitope (Ignatowicz et al., 1996). We can therefore view the process of negative selection as an attempt to reduce the probability of reacting to self to an acceptably low level. In agreement with this our results show that, whether division is included or not, encounter based deletion can never entirely be 100% efficient. However, previous work has argued this point on the basis of specific parameter values (Muller and Bonhoeffer, 2003). In general, when reviewing the literature in this area it has been difficult to visualize the effect of changes in parameter regimes. The production of contour plots in α and k parameter space addresses this issue.

These figures enable the assessment of the effect of negative selection over the entire range of α for any range of k. If we consider the physiologically relevant values of k, we see that these figures have proved particularly useful in comparing the 2 models under examination here. Since, analytic results in the limit $k \to \infty$ may have been of some value in determining the underlying differences in the models, but tell us little about their behaviour at these relatively low numbers of k.

6.5.2 Division Broadens the T-Cell Repertoire

Overall it appears that division would have the effect of broadening the repertoire of clones that survive negative selection on α. It does so because it increases insensitivity to k as our contour plots show. Our results suggest that this broadening would feature an increase in relatively high α clones and these would be present in low copy numbers. An immediate reaction to this would be that increasing the number of surviving high α clones would increase the prospect of auto-reactivity in the periphery. However, this issue is complicated in that auto-reactivity is not merely dependent

on α. This has consequences for how we examine the difference between the two models.

6.5.3 Frequency versus Numbers

Examination of the difference in the dividing and non-dividing cases uses 2 different measures: i) thymocyte frequency distributions and ii) thymocyte number distributions. These 2 measures may both have meaning within the context of auto-immunity. This is based on the simple notion that, in the periphery, the probability of a T cell auto-reacting depends on the probability of it having an encounter with an APC in the first place.

A simple illustration of this idea can be provided if we imagine a simple immune system with 100 T cells and 1 APC. We also imagine that included in our 100 T cells there are 5 cells which have an identical probability α of auto-reacting per encounter. Let us further imagine a discrete time frame, where each time step was equal to the duration of an encounter. We also assume that the number of encounters that the APC can accomodate at any one time with any T cell is 1. Given this scenario the probability of one of the 5 cells autoreacting per time step (P_{react}) is

$$P_{react} = \alpha \frac{5}{100}$$

The second term on the RHS of this equation is the frequency of the 5 cells in the total population and translates as the probability that one of our 5 cells will have an encounter. In this situation, where APCs interactions are limiting, the probability of auto-reaction therefore depends upon T cell frequency.

Imagine now that we are able to adjust the number of possible encounters per time step so that all our 5 cells can simultaneously have an encounter each. This can be achieved by i) increasing the number of APCs, ii) increasing the number of encounters an APC can accommodate or iii) a combination of both i and ii. In this scenario where APC interactions are non-limiting we find that,

$$P_{react} = \alpha$$

since the probability of having an encounter at each time step is now 1.

The expected number of auto-reactions for our 5 cells per time step N_{react} is given by

$$N_{react} = 5 \times P_{react}$$

In our first scenario N_{react} is therefore dependent on the frequency of our 5 cells. However, in our second non-limiting scenario N_{react} is dependent on their number, since in this case $P_{react} = \alpha$.

To sum up, if all T-cells with a common probability of auto-reacting per encounter can freely interact with APCs ie. APC interactions are not limiting, then the rate of auto-reactivity is dependent on T-cell number. In this context we therefore look to the thymocyte number distributions in order to distinguish the difference in the effects of the 2 cases examined here. However, if APC interactions are limiting then the rate of auto-reactivity is dependent on T-cell frequency. In this context, thymocyte frequency distributions are therefore our chosen means of differentiating between the models.

6.5.4 The Model Differences

Given we have 2 alternative contexts in which to view our results ie. APC interactions are or are not limiting, we now examine our results in relation to each.

Non-Limiting APC Interactions

Within the context of unlimited APC access the rate of auto-reactivity is dependent on T-cell numbers. Examining, our results for thymocyte numbers produced by non-dividing selection we note that number is always reduced and continues to be so with increasing k. Non-dividing selection can therefore be seen to lack robustness to k (see below).

In the case of dividing selection, we also note that in dividing selection for thymocytes with $\alpha < \gamma$ the population is always increasing. However, the point of intersection at $\alpha = \gamma$ where $E[Z_k] = 1$ for all k is of interest here in that if $\alpha > \gamma$ then the number of thymocytes is reduced. Speculatively, this point of intersection could act as a threshold which delineates the unacceptably highly auto-reactive from safer thymocytes. It would appear that division could act to enhance the difference between these 2 sub-populations and the difference is increased with increasing k.

Non-dividing selection has no such natural boundary and reduces the number of cells in both sub-populations. This suggests that non-dividing selection could act to reduce the highly auto-reactive population at the expense of the safe one. It seems obvious that reproductive success of an organism

may depend on the ability of it's immune system to fight pathogens. However, loss of reproductive success due to auto-immune disease is possibly equally damaging (Ghazeeri and Kutteh, 2001). Given the balance between these two factors, evolutionary pressure may therefore act to create an optimal tradeoff in terms of producing adequate numbers of safe T cells whilst attempting to avoid auto-immunity. If the pressure of providing adequate numbers is strong one possible limiting effect negative selection may have is to limit number of genes we possess (George, 2002). In summary, which of our models provides the "better deal" remains an open question. This is largely because of the difficulty in defining a measure of the relative value of safe versus unsafe T cells.

Limiting APC Interactions

In the context of limiting APC interactions our preliminary results suggest that rate of auto-reactivity depends on the frequency of thymocytes. In our results we find that there is, paradoxically perhaps, only a marginal difference between the dividing and non-dividing case in respect to this measure. This would suggest that division would allow us to increase the number of clones that survive negative selection whilst not adding substantially to the rate of auto-reactivity. Cross-reactivity in specificity means that we can view all clones as having some potential worth to the immune system, since even if they react strongly to self they may also react with a pathogen (Mason, 1998; Holler et al., 2003). Indeed, some pathogens imitate the host in order in to circumvent the immune system (Horwitz and Sarvetnick, 1999; Vogel et al., 2002). We hypothesize that it is possible to make a gain in clonal diversity, that is an increase the potential defensive armoury, via dividing selection and this could be traded off against the very small increase in the rate of auto-reactivity.

6.5.5 Non-dividing Selection is Not Robust

We have previously mentioned that in relation to thymocyte numbers non-dividing selection lacks robustness to the number of encounters. These distributions appear to drift with increasing k. Here we point out that this lack of robustness also extends to our results on the frequency of clones where the distributions continue to be altered substantially by increasing k. By way of contrast we see that this is not the case in the clonal results for dividing selection. These results show the posterior distributions become limiting for identical parameter regimes. However, the direction of the "drift" associated with non-dividing selection can be interpreted as being in its favour. In particular, if the reduction of P_k for higher α clones or thymocytes is the priority then, as a lower bound, non-dividing selection cannot be bettered.

One caveat to the reduction in thymocyte numbers caused by non-dividing selection is the proposal

that following selection some of the remaining thymocytes undergo expansion. Indeed, it has been suggested that the rate of expansion correlates with α (Le Campion et al., 2002). It is hoped that future work will examine the effect of this proposed regime of expansion. We therefore restrict ourselves to the comment that correlating expansion with α may probably be a dangerous tactic; elevating the frequency and number of those cells most likely to auto-react.

6.6 Conclusions

In summary we take from the above the following conclusions.

1. Division can broaden the T-cell repertoire by increasing the number of clonal specificities that survive selection. It does so by increasing the likelihood of survival for all clones. This can be attributed to the insensitivity to the number of encounters brought about by division.

2. If APCs are limited then the above may only have a small effect on the probability of auto-reactivity in comparison to non-dividing selection. We therefore hypothesize that division may present us with an advantageous tradeoff, allowing a broader range of clonal specificities whilst only incurring a very small cost in terms of increased likelihood of auto-reactivity.

3. If APCs are not limiting we note that the point intersection of the number distributions serves as a threshold above which thymocytes are reduced in numbers whilst below it they are increased. We therefore speculatively suggest that this may be a natural boundary between safe (below threshold) and unsafe (above threshold) cells.

4. Non-dividing selection is not robust to increasing encounters whilst due to the aforementioned insensitivity to encounter number dividing selection is.

Chapter 7

Conclusions

7.1 Review

The use of branching processes to model cell division in general is well established (Harris, 1963; Jagers, 1975). In particular, multitype branching processes have been successfully used to model epidemiological behaviour (Taneyhill et al., 1999; Kimmel and Axelrod, 2002). Here we have used the multitype process to model the evolution of CFSE dye in a labelled population of cells. Ultimately this has enabled us to draw biological conclusions from relatively simple models that either directly describe data (chapters 4 and 5) or make predictions about the effect of the hypothesis that selection and division may be concurrent (chapter 6). The work here is unique in concentrating on cell division and death in the thymus. However, the ability of our methods to model CFSE data and test hypotheses are not limited to the realm of immunology. The modelling approach could probably be transferred to investigate the behaviour of any dividing populations of cells.

When modelling biological systems it is not always possible or even desirable to model every detail. The process of abstraction often leads to the criticism that not all relevant information is included in a model. For example, the modelling contained within chapters 4 and 5 assumes that our thymocytes belong to a homogenous population; all thymocytes follow the same rules. The work in chapter 6 attempts to address this assumption by allowing for the fact that our probabilities of death, usually associated with thymocyte affinity for self-peptide/MHC complex, for a population of thymocytes is distributed in some way. To some extent this is similar to the modelling of T cell activation in the periphery (Gett and Hodgkin, 2000; Deenick et al., 2003; Leon et al., 2004; Hodgkin, 2005) where the probability of activation is stochastically modelled by assuming that the time to first division is normally distributed. Another approach is to adopt a more complex model of the cell cycle itself such as the Smith-Martin or the similar G_0 models (Bernard et al., 2003;

Pilyugin et al., 2003).

Throughout this thesis we have assumed that the probability of division is a constant for all thymocytes. It has been suggested that this is not the case (Le Campion et al., 2002). This later assumption may therefore be incorrect and further modelling could investigate the effect relaxing this constraint. This may be possible through the use of more general Bellman-Harris type processes of which the Markow age dependent (chapter 5) and Galton-Watson (chapter 4) are special cases. Caution is required, however, since the addition of further layers of complexity may render the results of the modelling intractable.

An alternative approach using discrete time branching processes is to create models with more than one population. For example the data we have examined could be described by a model that features two populations: one susceptible to death and one prone to division. However such models produce multiply redundant estimates on limited data such as presented by Hare et al. (1998) (personal observation). This issue can be resolved with more data. However, we point out that such limitations do not apply when these models are used as a purely exploratory or predictive tool.

As noted in section 3.2 the multinomial approximation is not easy to justify in that it is based on the assumption that each division category is independent of all others. Since each generation of a branching process is dependent upon the previous generation this cannot be strictly true. Answering the question as to why the multinomial approximation is able to produce reasonable estimates given that one of its primary assumptions is essentially incorrect remains a mathematical challenge.

7.2 Suggestions for further work

We have recently observed that isolated dead cells from CFSE labelled cell populations also display the distinct peaks associated with cell divisions. The data therefore indicated how death and division relate to each other in terms of how many cells die after undergoing a certain number of divisions. Given a population of CFSE stained cells, it therefore should possible to create a model that is able to combine this type of data and that derived from the usual analysis of living cells. A model would have to take into account the nature of the experimental technique. For example, in the thymus dead thymocytes are subject to clearance and this would require consideration. However, given that such difficulties can be accommodated this type of model may yield more useful information.

The methods relating to analysis of CFSE data described within this thesis are dependent upon

the data provided. In order to produce cell counts from an initial fluorescence histogram Hare et al. (1998) used a gating method (see figure 2. Hare et al. (1998)). This approach probably results in some cells being wrongly categorized. Such over-dispersion of data may be dealt with in several ways. The original data could be re-analysed with more sophisticated curve fitting techniques. For example, Gett and Hodgkin (2000) fitted each individual peak in a CFSE profile using a log-normal curve. An alternative approach, using branching processes, could attempt to directly model the probability of making an error in categorization.

Yet another alternative method which may refine the analysis of CFSE data is to construct a model which more fully represents the loss of fluorescence in dividing cells. This would take into account two things. Firstly, cells in a labelled population do not all contain identical quatities of dye as is evident from the distribution of fluorescence seen in a non-dividing sample (Figure 2. Hare et al. (1998)). In addition, the partition of dye between daughter cells may not be an exact 50% split. This means that as the number of divisions increases so the variance of each fluorescence peak will also increase. A model of this process could be produced by modelling the distribution of dye in the initial population as either a continuous or discrete probability distribution function. Subsequent stochastic modelling of division could take into account the uneven distribution of dye between daughter cells possibly also based on a distribution function.

In this thesis we also noted that the FTOC cultures we have analysed did not include DCs. This is thought to have a dramatic effect on the degree of negative selection that takes place (Anderson et al., 1998). It would be interesting to compare our results with those derived from cultures that contain DCs. The introduction of DCs may result in a number of outcomes. For example death may occur at the DP CD69$^+$ stage or at both this and the SP stages. One useful result of DC addition would be that parameter values would be more relevant to the thymus *in vivo*. This would allow comparisons to relevant data that either exists or may become available in the future. An outstanding possibility that could not be resolved here is that negative selection could be taking place prior to CD69 expression. Baldwin et al. (2005) argue that the timing of negative selection is dependent on the timing of TCRα chain expression in conjunction with the avidity for and localization of antigen. The system employed to arrive at these conclusions involved the timed expression of the TCRα chain. Modelling of the type used within this thesis could be used in conjunction with CFSE labelling and the timed expression of TCRα to produce a system to further explore such hypotheses.

Bibliography

K. Akashi, M. Kondo, and I. L. Weissman. Role of interleukin-7 in T-cell development from hematopoietic stem cells. *Immunol Rev*, 165:13–28, Oct 1998.

G. Anderson and E. J. Jenkinson. Lymphostromal interactions in thymic development and function. *Nat Rev Immunol*, 1(1):31–40, 2001.

G. Anderson, J. J. Owen, N. C. Moore, and E. J. Jenkinson. Thymic epithelial cells provide unique signals for positive selection of CD4+CD8+ thymocytes in vitro. *J Exp Med*, 179(6):2027–2031, Jun 1994.

G. Anderson, K. J. Hare, N. Platt, and E. J. Jenkinson. Discrimination between maintenance- and differentiation-inducing signals during initial and intermediate stages of positive selection. *Eur J Immunol*, 27(8):1838–42, 1997.

G. Anderson, K. M. Partington, and E. J. Jenkinson. Differential effects of peptide diversity and stromal cell type in positive and negative selection in the thymus. *J Immunol*, 161(12):6599–603, 1998.

G. Anderson, B. C. Harman, K. J. Hare, and E. J. Jenkinson. Microenvironmental regulation of t cell development in the thymus. *Semin Immunol*, 12(5):457–64, 2000.

Eric H Baehrecke. How death shapes life during development. *Nat Rev Mol Cell Biol*, 3(10):779–87, Oct 2002.

K. K. Baldwin, B. P. Trenchak, J. D. Altman, and M. M. Davis. Negative selection of t cells occurs throughout thymic development. *J Immunol*, 163(2):689–98, 1999.

Troy A Baldwin, Michelle M Sandau, Stephen C Jameson, and Kristin A Hogquist. The timing of TCR alpha expression critically influences T cell development and selection. *J Exp Med*, 202(1): 111–121, Jul 2005.

Stefan Beissert, Agatha Schwarz, and Thomas Schwarz. Regulatory T cells. *J Invest Dermatol*, 126(1):15–24, Jan 2006.

Samuel Bernard, Laurent Pujo-Menjouet, and Michael C Mackey. Analysis of cell kinetics using a cell division marker: mathematical modeling of experimental data. *Biophys J*, 84(5):3414–24, May 2003.

R. J. De Boer and A. S. Perelson. Estimating division and death rates from cfse data. *Journal of Computational and Applied Mathematics*, 184(1):140–164, 2005.

J. A. Borghans, A. J. Noest, and R. J. De Boer. How specific should immunological memory be? *J Immunol*, 163(2):569–75, 1999.

C. Bouneaud, P. Kourilsky, and P. Bousso. Impact of negative selection on the t cell repertoire reactive to a self-peptide: a large fraction of t cell clones escapes clonal deletion. *Immunity*, 13 (6):829–40, 2000.

R. Ceredig. Intrathymic proliferation of perinatal mouse alpha beta and gamma delta t cell receptor-expressing mature t cells. *Int Immunol*, 2(9):859–67, 1990.

C. Chan, R. Callard, O. Garden, AJ. George, S. Moon, F. Rovis, J. Stark, J. Strid, and A. Yates. Using cfse data to model heterogeneous cell populations. ., manuscript in preparation.

Eun Young Choi, Kyeong Cheon Jung, Hyo Jin Park, Doo Hyun Chung, Jin Sook Song, Seung Don Yang, Elizabeth Simpson, and Seong Hoe Park. Thymocyte-thymocyte interaction for efficient positive selection and maturation of CD4 T cells. *Immunity*, 23(4):387–396, Oct 2005.

Elissa K Deenick, Amanda V Gett, and Philip D Hodgkin. Stochastic model of T cell proliferation: a calculus revealing IL-2 regulation of precursor frequencies, cell cycle time, and survival. *J Immunol*, 170(10):4963–4972, May 2003.

B Efron. *An introduction to the bootstrap*. Chapman & Hall, 1993.

B. Ernst, C. D. Surh, and J. Sprent. Thymic selection and cell division. *J Exp Med*, 182(4):961–71, 1995.

M. J. Gabor, D. I. Godfrey, and R. Scollay. Recent thymic emigrants are distinct from most medullary thymocytes. *Eur J Immunol*, 27(8):2010–5, 1997.

Vitaly V Ganusov, Sergei S Pilyugin, Rob J de Boer, Kaja Murali-Krishna, Rafi Ahmed, and Rustom Antia. Quantifying cell turnover using CFSE data. *J Immunol Methods*, 298(1-2): 183–200, Mar 2005.

K Christopher Garcia and Erin J Adams. How the T cell receptor sees antigen–a structural view. *Cell*, 122(3):333–6, Aug 2005.

A. J. George. Is the number of genes we possess limited by the presence of an adaptive immune system? *Trends Immunol*, 23(7):351–5, 2002.

AV Gett and PD Hodgkin. A cellular calculus for signal integration by T cells. *Nat Immunol*, 1 (3):239–44, Sep 2000.

G. S. Ghazeeri and W. H. Kutteh. Autoimmune factors in reproductive failure. *Curr Opin Obstet Gynecol*, 13(3):287–91, 2001.

K. Hardy, S. Spanos, D. Becker, P. Iannelli, R. M. Winston, and J. Stark. From cell death to embryo arrest: mathematical models of human preimplantation embryo development. *Proc Natl Acad Sci U S A*, 98(4):1655–60, 2001.

K. J. Hare, R. W. Wilkinson, E. J. Jenkinson, and G. Anderson. Identification of a developmentally regulated phase of postselection expansion driven by thymic epithelium. *J Immunol*, 160(8): 3666–72, 1998.

K. J. Hare, E. J. Jenkinson, and G. Anderson. Cd69 expression discriminates mhc-dependent and -independent stages of thymocyte positive selection. *J Immunol*, 162(7):3978–83, 1999a.

K. J. Hare, E. J. Jenkinson, and G. Anderson. In vitro models of t cell development. *Semin Immunol*, 11(1):3–12, 1999b.

K. J. Hare, E. J. Jenkinson, and G. Anderson. An essential role for the il-7 receptor during intrathymic expansion of the positively selected neonatal t cell repertoire. *J Immunol*, 165(5): 2410–4, 2000.

TE Harris. *The Theory of Branching Processes*. Springer-Verlag, Berlin, 1963.

PD Hodgkin. Quantitative rules for lymphocyte regulation: the cellular calculus and decisions between tolerance and activation. *Tissue Antigens*, 66(4):259–66, Oct 2005.

Kristin A Hogquist, Troy A Baldwin, and Stephen C Jameson. Central tolerance: learning self-control in the thymus. *Nat Rev Immunol*, 5(10):772–82, Oct 2005.

P. D. Holler, L. K. Chlewicki, and D. M. Kranz. Tcrs with high affinity for foreign pmhc show self-reactivity. *Nat Immunol*, 4(1):55–62, 2003.

M. S. Horwitz and N. Sarvetnick. Viruses, host responses, and autoimmunity. *Immunol Rev*, 169: 241–53, 1999.

AL Hughes and MK Hughes. Self peptides bound by HLA class I molecules are derived from highly conserved regions of a set of evolutionarily conserved proteins. *Immunogenetics*, 41(5):257–62, 1995.

L. Ignatowicz, J. Kappler, and P. Marrack. The repertoire of t cells shaped by a single mhc/peptide ligand. *Cell*, 84(4):521–9, 1996.

P Jagers. *Branching Processes with Biological Applications*. Wiley Series in Probability and Mathematical Statistics-Applied. John Wiley & Sons, London, 1975.

SC Jameson, KA Hogquist, and MJ Bevan. Positive selection of thymocytes. *Annu Rev Immunol*, 13:93–126, 1995.

CA. Janeway, P. Travers, M. Walport, and MJ. Shlomchik. *Immunobiology: The immune system in health and disease.* Garland Publishing, 5th edition, 2001.

Kimmel and Axelrod. *Branching Processes in Biology.* Springer-Verlag, New York, 2002.

H. Kishimoto and J. Sprent. Negative selection in the thymus includes semimature t cells. *J Exp Med*, 185(2):263–71, 1997.

H. Kishimoto, C. D. Surh, and J. Sprent. A role for fas in negative selection of thymocytes in vivo. *J Exp Med*, 187(9):1427–38, 1998.

P. Kisielow, H. Blthmann, U. D. Staerz, M. Steinmetz, and H. von Boehmer. Tolerance in T-cell-receptor transgenic mice involves deletion of nonmature CD4+8+ thymocytes. *Nature*, 333 (6175):742–746, Jun 1988.

Mitchell Kronenberg and Alexander Rudensky. Regulation of immunity by self-reactive T cells. *Nature*, 435(7042):598–604, Jun 2005.

T. M. Laufer, J. DeKoning, J. S. Markowitz, D. Lo, and L. H. Glimcher. Unopposed positive selection and autoreactivity in mice expressing class ii mhc only on thymic cortex. *Nature*, 383 (6595):81–5, 1996.

TM Laufer, LH Glimcher, and D Lo. Using thymus anatomy to dissect T cell repertoire selection. *Semin Immunol*, 11(1):65–70, Feb 1999.

Alfons Lawen. Apoptosis-an introduction. *Bioessays*, 25(9):888–96, Sep 2003.

A. Le Campion, F. Vasseur, and C. Penit. Regulation and kinetics of premigrant thymocyte expansion. *Eur J Immunol*, 30(3):738–46, 2000.

A. Le Campion, B. Lucas, N. Dautigny, S. Leaument, F. Vasseur, and C. Penit. Quantitative and qualitative adjustment of thymic t cell production by clonal expansion of premigrant thymocytes. *J Immunol*, 168(4):1664–71, 2002.

R Lechler, JG Chai, F Marelli-Berg, and G Lombardi. T-cell anergy and peripheral T-cell tolerance. *Philos Trans R Soc Lond B Biol Sci*, 356(1409):625–37, May 2001.

Kalet Leon, Jose Faro, and Jorge Carneiro. A general mathematical framework to model generation structure in a population of asynchronously dividing cells. *J Theor Biol*, 229(4):455–76, Aug 2004.

Little and Rubin. *Statistical analysis with missing data.* Probability and mathematical statistics. Wiley, New York, 1987.

AB Lyons. Divided we stand: tracking cell proliferation with carboxyfluorescein diacetate succinimidyl ester. *Immunol Cell Biol*, 77(6):509–15, Dec 1999.

AB Lyons. Analysing cell division in vivo and in vitro using flow cytometric measurement of CFSE dye dilution. *J Immunol Methods*, 243(1-2):147–54, Sep 2000.

D. Mason. A very high level of crossreactivity is an essential feature of the t-cell receptor. *Immunol Today*, 19(9):395–404, 1998.

R Mehr, A Globerson, and AS Perelson. Modeling positive and negative selection and differentiation processes in the thymus. *J Theor Biol*, 175(1):103–26, Jul 1995.

F.A. Mood, A.M. & Graybill. *Introduction to the theory of statistics*. McGraw-Hill Book Company, international student edition second edition edition, 1963.

V. Muller and S. Bonhoeffer. Quantitative constraints on the scope of negative selection. *Trends Immunol*, 24(3):132–5, 2003.

GJ Nossal. Negative selection of lymphocytes. *Cell*, 76(2):229–39, Jan 1994.

E. Palmer. Negative selection–clearing out the bad apples from the t-cell repertoire. *Nat Rev Immunol*, 3(5):383–91, 2003.

C. Penit and F. Vasseur. Expansion of mature thymocyte subsets before emigration to the periphery. *J Immunol*, 159(10):4848–56, 1997.

Sergei S Pilyugin, Vitaly V Ganusov, Kaja Murali-Krishna, Rafi Ahmed, and Rustom Antia. The rescaling method for quantifying the turnover of cell populations. *J Theor Biol*, 225(2):275–83, Nov 2003.

WH. Press, SA. Teukolsky, WT. Vetterling, and BP. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, second edition, 2002.

M. Regner. Cross-reactivity in t-cell antigen recognition. *Immunol Cell Biol*, 79(2):91–100, 2001.

Derek B Sant'Angelo and Charles A Janeway. Negative selection of thymocytes expressing the D10 TCR. *Proc Natl Acad Sci U S A*, 99(10):6931–6936, May 2002.

R. Scollay and D. I. Godfrey. Thymic emigration: conveyor belts or lucky dips? *Immunol Today*, 16(6):268–73; discussion 273–4, 1995.

E. Sebzda, S. Mariathasan, T. Ohteki, R. Jones, M. F. Bachmann, and P. S. Ohashi. Selection of the t cell repertoire. *Annu Rev Immunol*, 17:829–74, 1999.

B. Seddon and D. Mason. The third function of the thymus. *Immunol Today*, 21(2):95–9, 2000.

C. L. Sentman, J. R. Shutter, D. Hockenbery, O. Kanagawa, and S. J. Korsmeyer. bcl-2 inhibits multiple forms of apoptosis but not negative selection in thymocytes. *Cell*, 67(5):879–888, Nov 1991.

Jonathan Sprent and Hidehiro Kishimoto. The thymus and negative selection. *Immunol Rev*, 185: 126–35, Jul 2002.

A.F. Stanton, R.E. Bleil, and S. Kais. A new approach to global minimization. *Journal of Computational Chemistry*, 18(4):594–599, 1997.

A. Strasser, A. W. Harris, H. von Boehmer, and S. Cory. Positive and negative selection of T cells in T-cell receptor transgenic mice expressing a bcl-2 transgene. *Proc Natl Acad Sci U S A*, 91 (4):1376–1380, Feb 1994.

DE Taneyhill, AM Dunn, and MJ Hatcher. The Galton-Watson branching process as a quantitative tool in parasitology. *Parasitol Today*, 15(4):159–65, Apr 1999.

H. A. Van Den Berg, D. A. Rand, and N. J. Burroughs. A reliable and safe t cell repertoire based on low-affinity t cell receptors. *J Theor Biol*, 209(4):465–86, 2001.

A. Vogel, M. P. Manns, and C. P. Strassburg. Autoimmunity and viruses. *Clin Liver Dis*, 6(3): 451–65, 2002.

H. von Boehmer. Developmental biology of t cells in t cell-receptor transgenic mice. *Annu Rev Immunol*, 8:531–56, 1990.

H. von Boehmer and H. J. Fehling. Structure and function of the pre-t cell receptor. *Annu Rev Immunol*, 15:433–52, 1997.

R. W. Wilkinson, G. Anderson, J. J. Owen, and E. J. Jenkinson. Positive selection of thymocytes involves sustained interactions with the thymic microenvironment. *J Immunol*, 155(11):5234–5240, Dec 1995.